LONDON'S GLOBAL UNIVERSITY
Department of Computer Science

Optimising HVAC Control Across Diverse Climates: A Replay-Enhanced Deep Reinforcement Learning Approach

Nathan Carey

Supervisor: John Shawe-Taylor

Co-Supervisors: Francesca Channon, Dhruva Tirumala

Faculty of Engineering

Department of Computer Science

University College London

A Project Report Presented in Partial Fulfillment of the Degree

MSc AI for Sustainable Development

September 2024

Executive Summary

Background

In this thesis, we address the inefficiencies of traditional rule-based control (RBC) in managing heating, ventilation, and air conditioning (HVAC) systems. These systems, which account for up to **40%** of total energy consumption in commercial buildings [1], behave nonlinearly, especially under varying weather conditions. This makes RBCs insufficient for optimisation in a modern, changing environment. We explore the potential of deep reinforcement learning (DRL) to provide more adaptive and efficient control strategies. However, DRL approaches thus far have significant limitations, including a lack of generalisation across different environments, high computational costs [2], and insufficient evaluation in real-world scenarios [3].

Research Question

How can we create a DRL agent capable of efficiently managing HVAC systems across diverse and rapidly changing climate conditions?

Research Contributions

- **Three Climate Experiments:** We developed a novel Three Climate Experiment framework to improve the robustness of DRL agents by training them across diverse climate conditions (seen in Figure 1). We conducted sequential experiments across different climates, allowing the DRL agent to leverage knowledge from previous

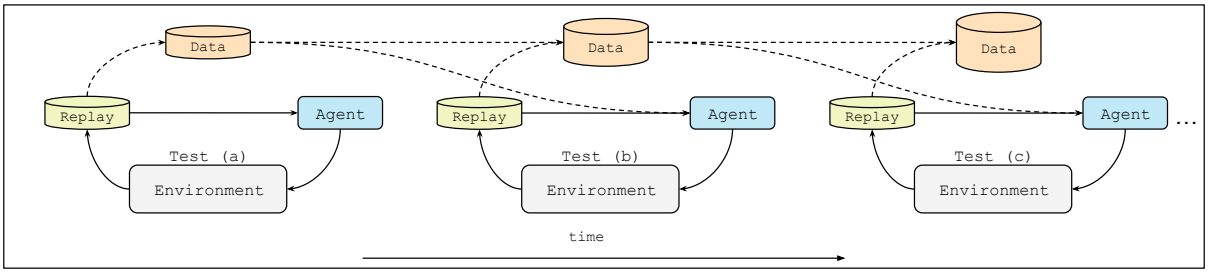


Figure 1: Three Climate Experiment Framework

scenarios. This incorporates the Replay across Experiments (RaE) methodology [4], improving stability and performance in both normal and extreme weather conditions.

- **Dataset Size Comparison:** We compared the agent's performance with two different experience levels: 10,000 and 100,000 sampled transitions, demonstrating how increased experience improves model robustness and effectiveness.
- **Novel Weather Dataset:** We introduced a novel dataset composed of global weather data from seven countries (Appendix Table A.2). Additionally, we modified this dataset to reflect current climate variability, providing a more realistic and challenging evaluation environment.

Methodology

We employed a structured approach, as seen in Figure 2. First, we reproduced and refactored a baseline DRL algorithm for HVAC [5]. Next, we implemented the Three Climate Experiments, training our DRL models across three distinct climate conditions, leveraging the RaE approach to build robustness by incorporating data from previous trials [4]. Performance was evaluated using both standard DRL metrics (e.g., average reward, mean squared error loss) and HVAC-specific metrics (e.g., energy consumption, thermal comfort violation) to ensure comprehensive assessment (Appendix Table A.4) Our novel, global weather dataset was modified to reflect current climate conditions by adding Gaussian noise and data perturbations (Appendix Figure A.3). This dataset was used to compare model robustness to extreme climate variations.

Key Results and Findings

Our research demonstrates that the Three Climate Experiment framework significantly improves upon the robustness and generalisation of traditional DRL agents in HVAC control. As summarized in Table 1, we trained across multiple climate conditions, and we

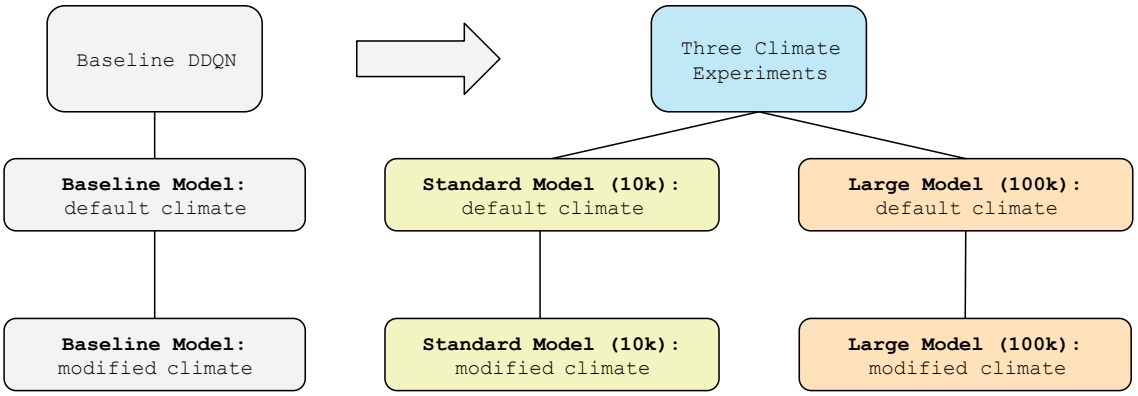


Figure 2: Overview of Experiments

observed up to a **53.02%** improvement in energy efficiency over the baseline approach. The RaE methodology further augmented the models' ability to adapt to new environments by incorporating prior experiences into its learning process, leading to faster convergence and improved performance in both typical and extreme weather scenarios. Our models were further validated with the modified climate dataset to better represent the impacts of climate change, ensuring testing against realistic conditions. In that challenging environment, they outperformed energy-saving baselines by an average of **39.057 percentage points**.

We hypothesise that the significant energy savings are primarily due to the identification of Zone 3 as a key area where optimised control yields the greatest overall efficiency for the entire building (as shown in Figure 3).

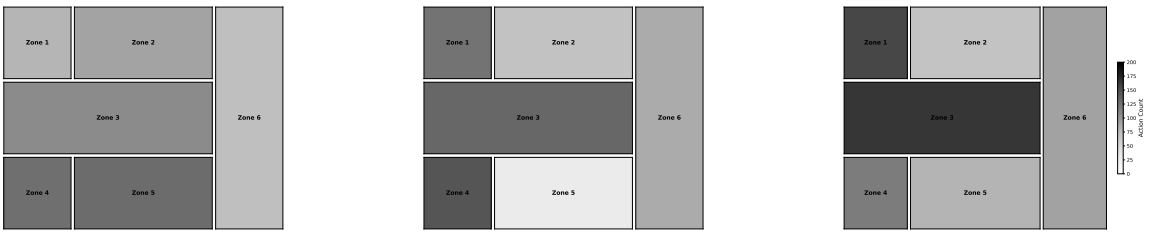
We validated our hypotheses with several control-level examples showcasing the relative smoothness of our solution (Figure 3.1 Appendix Figure A.5), and engaged with building managers to explain these results in Section B. These efforts toward better understanding the models' decisions are necessary steps in preparing for the online deployment of these models in the real world, and making an impact on the United Nations Sustainable Development Goals (SDGs).

Category	Performance	Description	Impact
Baseline	Reproduced baseline within +/- 6.4% accuracy	Established solid foundation for comparison across models.	Published refactored code as fully open-source
Three Climate (10k)	**13.34%** improvement over baseline (max)	Improvement shown in energy efficiency. However, performance suffers in desert climates (AZ, Dubai).	**162,014 kWh p.a.** energy savings
Three Climate (100k)	**53.02%** improvement over baseline (max)	Significant gains in performance and stability achieved with the larger model.	**378,458 kWh p.a.** energy savings
Modified Weather (10k, 100k)	**54.11%** improvement in challenging climates (max)	The Three Climate Experiment models showed robustness against climate variability. The 100k model outperformed baselines in all cases.	**523,616 kWh p.a.** energy savings
Total Impact: 1.064 million kWh p.a. energy savings			

Table 1: Impact Summary of Key Contributions

This thesis contributes to the field of HVAC optimisation by providing a robust, efficient, and explainable DRL framework capable of optimising HVAC systems across diverse climates. The Three Climate Experiment methodology represents significant advancements in the generalisation and efficiency of DRL models for HVAC control.

Future work should focus on extending this research to a real building, with the potential of incorporating real-world factors (such as energy pricing) and exploring more advanced DRL models. Further efforts should be made towards the exhaustive explainability of these models, ensuring that they can be effectively implemented and trusted in real-world applications.



(a) Baseline DDQN (b) Three Climate (c) Three Climate (Large)

Figure 3: Comparison of Action Distributions for Agents in Vancouver (modified climate). Darker Zones Represent More Counts of Enforcing Temperature Control.

In conclusion, we have demonstrated that a replay-enhanced DRL approach can significantly improve HVAC control, achieving substantial energy savings while maintaining occupant comfort. The introduction of the Three Climate Experiment methodology marks a step forward in the development of generalisable DRL models for HVAC optimisation and directly addresses several SDGs, particularly SDG 7 (Affordable and Clean Energy), SDG 11 (Sustainable Cities and Communities), and SDG 13 (Climate Action) [6].

Abstract

Heating, ventilation, and air conditioning (HVAC) systems present a major challenge to energy efficiency in the built environment. These systems consume as much as half of the total energy, and their behavior is complex and nonlinear. This makes traditional rule-based control (RBC) for optimization inefficient. Reinforcement Learning (RL) and Deep Reinforcement Learning (DRL) have emerged as potential solutions, with growing research interest and applications in simulation environments. However, most research is based on models that struggle to generalise across rapidly changing weather patterns and are ill-equipped for real-world control. Additionally, many promising DRL models for HVAC controls have been trained with an enormous amount of computation, evaluated on statistically insignificant periods of time, or rely on pseudo-actions that do not correspond to real-world controls.

To address these limitations, we introduce an effective framework for a more capable agent in multi-zone HVAC optimisation. Central to this framework is the Three Climate Experiment, a novel approach designed to improve agent robustness by training across diverse climate conditions in sequential experiments. This framework incorporates the "Replay across Experiments" (RaE) methodology [4], enabling the agent to leverage learned experiences from previous scenarios to stabilise performance in both normal and extreme weather conditions. We present empirical evidence showing that our approach not only enhances robustness across sensitive weather environments, outperforming previous methods by up to **54.11%** in energy efficiency but also improves data efficiency by allowing the model to converge within a single epoch of training. Finally, we address the explainability concerns towards online applications and the steps necessary for transitioning from simulated environments to real-world control systems.

Keywords— Reinforcement Learning, HVAC, Experience Replay, Energy Efficiency, Multi-zone Optimisation

All models are wrong, but some are useful.

— George E.P. Box

Acknowledgements

I would like to express my gratitude to my supervisors, **Professor John Shawe-Taylor** and **Francesca Channon** whose expertise, support, and constant encouragement helped me throughout the research process.

Special thanks to **Naqash Tahir** and **PGIM** for their collaboration and valuable insights, inspiring this project and informing the practical implications of this work. This is instrumental in shaping the direction of my research.

I would also like to thank **Eduardo Pignatelli** and **Pandarasamy Arjunan** for their support in my submission to the Tackling Climate Change with Machine Learning Workshop at NeurIPS 2024.

I would like to thank my colleagues at University College London: Björn, Ilai, Lukas, Mert, Naomi, and Shahar for their stimulating discussions and collaborative spirit. I have learned much from each of you.

Finally, I would like to thank the researchers in the Clemson AIS lab for their open-source contributions and guidance, which helped me gain a strong foothold in tackling this problem.

Nathan Carey

Declaration

I, Nathan Carey, declare that the thesis has been composed by myself and that the work has not be submitted for any other degree or professional qualification. I confirm that the work submitted is my own, except where work which has formed part of jointly-authored publications has been included and referenced. The report may be freely copied and distributed provided the source is explicitly acknowledged.

09/09/24

Signature *Date*

Table of Contents

List of Figures

List of Tables

List of Algorithms

1 | Introduction and Background

1.1 Introduction

The impacts of climate change have become increasingly apparent and severe in recent years [7]. Extreme weather events, including storms, prolonged droughts, devastating wildfires, and severe flooding, are now occurring with alarming frequency and intensity. These changes are rapidly reshaping global ecosystems, threatening biodiversity, and putting significant strain on the natural resources and agricultural systems that are crucial for human survival.

In 2023, we witnessed the **hottest year on record**, surpassing previous temperature highs by a considerable margin. The trend of record-breaking temperatures has continued into 2024, with January marking another month of unprecedented warmth. According to data from the Copernicus Climate Change Service, January 2024 was the warmest January ever recorded globally (Figure 1.1). This continues the pattern of exceptional heat observed in recent years, underscoring the accelerating pace of global warming.

Current projections suggest that without immediate and drastic action to eliminate global greenhouse gas emissions, we face catastrophic consequences within the coming decades. A 2024 forecast published in Nature predicts that by 2050, climate change could cause average incomes to fall by almost 20% and result in $38 trillion of economic destruction annually [8]. Despite these dire warnings and the increasing visibility of climate impacts, global emissions continue to rise year after year. The difference in emissions between necessary carbon dioxide reduction and current trends is estimated at 12-15 gigatons of CO2-equivalent if we are to limit global warming to 1.5°C [9]. This persistent increase in emissions, coupled with the intensifying effects of climate change, has created a feedback loop that accelerates the climate crisis at an unprecedented rate [7, 10].

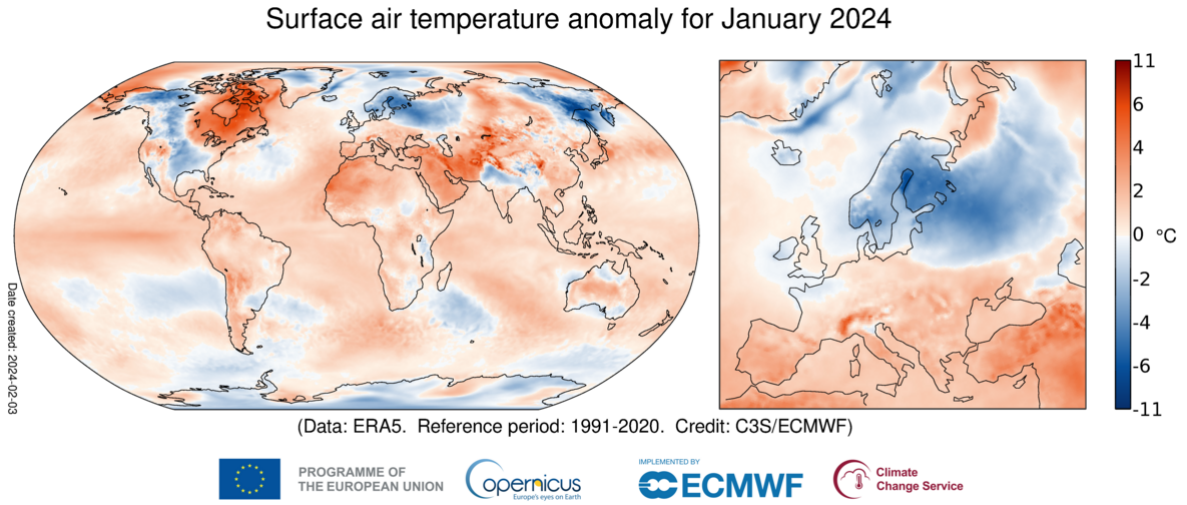


Figure 1.1: Surface Air Temperature Anomaly for January 2024 Relative to the January Average for the Period 1991-2020.

These climate trends have a significant impact on the outdoor weather thermodynamics that affect building energy consumption, particularly in HVAC systems. As global temperatures rise and extreme weather events become more frequent, the demand for heating and cooling in buildings is expected to increase dramatically. This underscores the critical need for more efficient HVAC systems and control strategies that can adapt to changing climate conditions while minimising energy consumption.

Our research aims to address this challenge, while working towards several United Nations Sustainable Development Goals (SDGs), particularly SDG 7 (Affordable and Clean Energy), SDG 11 (Sustainable Cities and Communities), and SDG 13 (Climate Action) [6]. Specifically, we seek to answer the question: how can we create a DRL agent capable of efficiently managing HVAC systems across diverse and rapidly changing climate conditions?

1.1.1 Impact of the built environment

The built environment is a major contributor to CO2 emissions globally. In 2022, buildings, including their construction, accounted for about 34% of global final energy consumption and 37% of global energy-related CO2 emissions. This includes both direct emissions from on-site energy use and indirect emissions from the production of electricity and heat used in buildings [1]. HVAC systems significantly impact building energy consumption, and they can account for up to **40%** of total energy used in commercial structures. Given the substantial energy footprint of HVAC systems in buildings, our research focuses on leveraging DRL techniques to optimise HVAC control. By developing more intelligent and adaptive control strategies through DRL, we aim to significantly reduce energy consumption while maintaining occupant comfort, thereby addressing a critical aspect of building sustainability [6].

1.1.2 Current approaches in building energy management

Building Energy Management Systems (BEMS) have been developed to improve energy efficiency in commercial buildings. These systems rely on feedback from various technologies, including sensors, data analysis tools, and control algorithms, to monitor and manage energy-consuming systems. Modern commercial buildings equipped with BEMS can use smart sensors to adjust energy consumption dynamically based on occupancy and other factors. These buildings' centralization of HVAC systems allows for more advanced control algorithms. BEMS provide valuable insights into energy usage patterns, identify energy-saving opportunities, and enable proactive energy management strategies. Currently, the three most proven approaches to energy control in BEMS are rule-based control (RBC), model-predictive control (MPC), and RL-based.

While each approach offers distinct advantages, DRL-based research is particularly compelling for HVAC optimisation due to its ability to enable BEMS to deliver a more

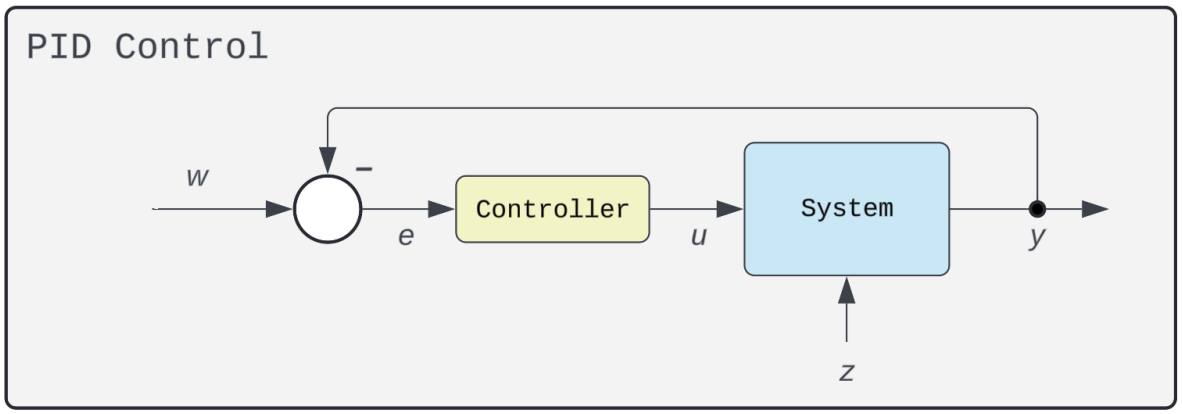


Figure 1.2: Classical Feedback Control Loop (PID)

adaptable and autonomous solution. Unlike traditional methods, DRL minimizes the need for specialised human intervention in the optimisation process, allowing building managers to allocate their attention to other critical aspects such as maintenance and subsystem upgrades. Our research seeks to advance this autonomy, bringing us closer to a future where buildings can adapt intelligently to varied environments.

(A) BEMS with Rule-Based Control (RBC):

In RBC, experts directly encode empirical data and prior knowledge to construct simple and reliable reactive control systems. This conventional approach is historically favored for its simplicity. Examples include adaptive occupancy-based lighting control with grey prediction models and cascade proportional integral derivative (PID) controllers for HVAC systems (Fig. 1.2). However, as commercial buildings become more complex, the inflexibility of rule-based strategies can lead to lower energy efficiency. Regardless of design methodology, RBC fails as it is limited by system dynamics, and it is difficult to find parameterizations in the nonlinear case [11, 12]. Significant energy waste occurs due to inadequate optimisation of unoccupied spaces, unnecessary thermal comfort maintenance during non-working hours, and inappropriate policies in areas like restrooms and storage facilities.

RBC systems may also fail to adjust for seasonal changes or special events, leading to excessive heating or cooling during off-hours. BEMS lose valuable energy performance when these inefficiencies are not addressed. Balancing energy optimisation with occupant health and comfort is also a challenging task that rule-based control systems struggle to address effectively. As building complexity increases, more sophisticated approaches are needed to achieve optimal energy efficiency while maintaining occupant well-being.

(B) Model Predictive Control-based BEMS (MPC):

Model Predictive Control (MPC) proves to be a sophisticated approach in the context of building energy management. MPC shifts the focus from the design of the controller to the design of a process model to predict future behavior of a controlled system. Figure 1.3 is a simple example of this.

MPC provides many improvements over RBC in the context of HVAC. Multiple variables can be easily considered such as temperature set point, flow rate set point, valve position, etc. MPC introduces feed-forward control to compensate for the measurable disturbance factors such as climatic predictions and occupancy rate. The objective is to minimize a cost function that can be calculated within well-defined constraints. Finally, there is a large group of techniques that can solve these optimisation problems: quadratic programming (QP), linear programming (LP), dynamic programming (DP), genetic algorithm (GA), particle swarm optimisation (PSO), etc.

Recent advancements in MPC applied to built environments have led to HVAC control strategies that demonstrate effective energy optimisation results [13, 14, 15, 16, 17]. The MPC methodology applied to built environments begins with developing a digital twin of the real building, which accurately replicates its structure and control mechanisms. This digital representation predicts temperature changes and analyses the thermodynamic behavior. By employing the digital twin, MPC can optimise control inputs, such as set-points and actuator functions, to meet energy efficiency and occupant comfort goals while

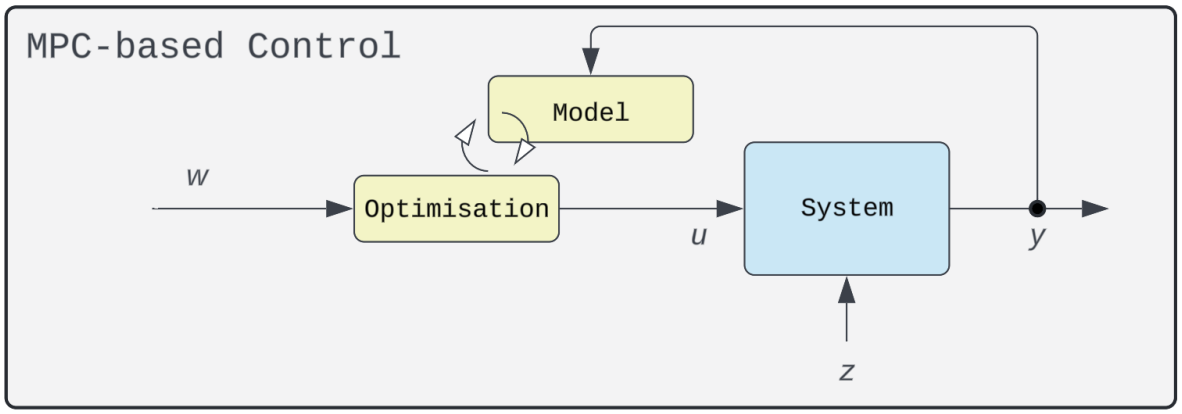


Figure 1.3: Simplified MPC-Based Control Loop

adhering to operational constraints. Additionally, optimisation-based building control, a subset of MPC, employs the aforementioned meta-heuristic optimisation algorithms like PSO to determine the best control inputs within the parameters set by the MPC framework.

Despite the effectiveness of MPC in managing building energy, HVAC systems operate in dynamic environments affected by various factors external to the deterministic control environment: including outdoor temperatures, solar radiation, occupancy trends, and internal heat gains. These factors can be challenging to incorporate into the MPC model due to their complex physical interactions. Furthermore, creating and maintaining an accurate digital model of a commercial building can be labor-intensive and time-consuming. Lastly, the lack of generalisability across environments is a concern as standard MPC defines a unique model for each building without considerations for learning across experiences or locations.

(C) Reinforcement Learning-based BEMS:

The third prominent approach in building energy management involves utilising RL and/or DRL. These machine learning-based methods enable agents to optimise energy consumption through data-driven modeling of the BEMS. These models continuously

improve control strategies based on state-action-reward mechanisms, learning through interaction with the BEMS. One additional benefit of RL is that it does not require expert knowledge of the physics involved in heat conduction or external influences, making it a flexible framework potentially requiring less specialised human knowledge to achieve results.

In the online case, the building or simulation engine is under direct control of an RL agent that is both learning and setting HVAC schedules in real-time. They can also be deployed offline, with the agent learning from a static dataset before deploying to control the building. In offline cases, the agent will not use the new experience to learn, in the neural network sense. Both approaches have gained considerable attention in research [18, 19].

There have also been considerable scientific breakthroughs with DRL which has opened up new capabilities for agents to solve this nuanced optimisation problem. Many deep RL algorithms have become increasingly popular in energy management. For example, a Deep Q-Network (DQN) with a memory buffer was implemented to manage airflow rates in various zones, demonstrating the effectiveness of DRL in addressing energy optimisation challenges in complex settings [18].

Zhang et al. employed a multi-agent solution, an enhanced DRL algorithm known as Asynchronous Advantage Actor Critic (A3C) in a real building to regulate supply water temperature. This method, when combined with MPC, achieved significant energy savings while ensuring occupant comfort [2].

Similarly, the Deep Deterministic Policy Gradient (DDPG) algorithm was applied to HVAC systems to overcome the limitations of DQN in continuous control scenarios [3]. Multi-Agent Deep Reinforcement Learning (MADRL) was also used to optimise building energy management by controlling supply air rates and damper positions, resulting in substantial energy savings in a commercial building with 30 zones [20].

Nguyen et al. show that Phasic Policy Gradients (PPG) outperform the Proximal Policy Optimisation (PPO). This significant advancement on the conventional method in HVAC control optimisation reduces energy consumption by 2-14% while enhancing indoor temperature comfort. They also achieve faster convergence while being less resource-intensive and easier to implement [21].

1.1.3 Challenges

DRL has emerged as the most flexible solution to the HVAC energy management problem, but there remain significant gaps in the state-of-the-art (SOTA) research.

DRL applications are rarely evaluated on real buildings. There is a significant dependence on simulation environments like EnergyPlus [15]. Some examples of this include [18, 22, 23, 3]. This is problematic as severe changes in the weather, HVAC system conditions, network outages, and other factors can degrade the accuracy of these simulations, even when strict adherence to building model size, internal components, and zone details is maintained [14].

DRL methods are highly parameterised and can take a significant amount of time to train, up to 10 hours in some cases. When over-complicated DRL models are deployed with a large number of parameters, they improve energy savings but increase training time. In real-world deployment, we expect these models to adapt to time-sensitive changes in the environment [24]. Online deployment will be unsuccessful with such methods [2, 3].

Most methods fail to employ prior knowledge and could be improved by introducing existing historical building management system (BMS) data to better equip the RL agents for real-world deployments. Although it is difficult to craft an offline environment with sufficient real-world data to provide a training environment, applications can introduce this data with more iterative methods. In contrast to prior work, we include data from past experiments inside of the working memory of a Double-Q Network agent (replay

buffer). Other applications that have attempted to address this have suffered from exceptionally high costs [24, 23]. We present a simple method of inserting this data into the replay buffer that leads to significantly higher performance gains without negatively impacting training times.

Most applications do not generalise well when tested across diverse environments and only perform well in-distribution of their training data. As climate change continues to worsen [7], this technical limitation becomes even more valuable to address. Our timely and reasonable goal is to provide solutions that reduce energy consumption and maintain a safe and comfortable indoor environment for all occupants that is effective across highly varying environments.

Current research may also rely on data that is insignificant or impossible to obtain in the real world such as room temperature, diffuse solar radiation, real-time occupancy [23], and wind direction. For example, literature review and engagement with practitioners have led us to understand that simple room-level temperature sensors are not common in BEMS systems for buildings equipped with ducts [25]. This is because the sensors that would measure temperature are usually at a junction of multiple zones. These sensors instead report a volume-based mix of zone temperatures. Therefore it is challenging to deploy a DRL model based on room temperature to this zone-based level of control over a BEMS.

Current approaches rarely explain where the energy savings originate or fail to compare their results to an appropriate baseline, over a significant amount of time [26]. We will attempt to address this in the explainability section of the discussion.

1.1.4 Our contributions

We categorise the contributions of this work into three main areas.

Baseline Reproduction and Algorithm Enhancement:

- We reproduced a DRL-based control algorithm to use as a baseline.
- We refactored the baseline algorithm to provide an object-oriented, modular approach that is fully open-sourced.
- We introduce a decaying exploratory parameter (epsilon).

Framework Innovation and Generalisation:

- We applied the novel Replay Across Experiments (RaE) workflow [4], resulting in improved performance across climates of up to **54%**, compared to baseline.
- We compare dataset sizes and data sources to relevant baselines (random weight resetting, base DDQN).
- We report results on a novel, open-sourced dataset comprising a globally representative sample of climates.

Efficiency, Adaptability, and Explainability:

- Our model requires only minimal input variables (outdoor temperature, indoor temperature, time, and control signals). The action space is a binary vector to activate/inactivate enforcing temperature range, instead of using explicit set points.
- Our model addresses the challenge of computational complexity by remaining efficient in both training and evaluation; the model takes seven minutes per epoch during training and evaluation (under sufficient hardware).
- We analysed real building data, engaged with building managers and industry professionals to explain our results, evaluate feasibility of real-wold deployment, and understand what future work is required.

The code for this project, including our control framework, datasets, and visualisations are open-source and available at https://github.com.

1.2 Related work

The use of Reinforcement Learning for HVAC control is an active research topic that has been studied for some time [27, 18, 19, 28]. HVAC controllers trained by RL algorithms have a consistent form: a mapping from the state space of a built environment to some action in the relevant HVAC subsystem. These actions could be enforceable temperature zones [5], temperature set points [29], lighting and blinds [30], or frequency setpoints of a pump or a cooling tower fan [31].

Recent research has focused on using a special type of RL called Q-learning [32, 33]. Q-learning is a type of value-based RL, where the goal is to learn towards an accurate value of each state and action. For example, the value of the agent being in a building with indoor temperature "x" taking action "y". This has been shown to achieve SOTA results when applied to individual problems in HVAC control [34, 31, 35, 28, 5].

These Q-learning controllers can control an entire building [5], a single zone [35], work together with other agents [31, 35], and/or operate on a specific piece of technology like a water tower or a chiller [31].

These agents can be further equipped to model uncertainty through Bayesian methods, although the study reviewed [35] only reports results from a 50-hour period. This is an insignificant amount of time, especially in regions with variable weather conditions across four seasons.

There is also a body of relevant research that extends an RL agent for various robotics tasks to share experience across experiments. This is typically done through reloading into the replay buffer during online learning [36, 37]. This insight does not require any changes with respect to the RL agent, and is generally agnostic to agent and architecture changes across experiments. An approach called AWAC extends this to combine fine-tuning offline agents with mixing data into a replay [38].

By definition, this solution requires multiple training stages [39, 40], but this approach creates a simple workflow that leads to excellent results from sequences of experiments that pass data along from training runs [4].

1.2.1 Relevant studies

The following relevant studies are sourced from a structured literature review process.

Deng, Zhang, Zhang, and Qi [35] propose a "non-stationary" DQN for optimal HVAC control using data from EnergyPlus and Building Controls Virtual Test Bed (BCVTB) simulation environments. They identify changing points in building environments claiming stability against disturbances and generalisation gains when applied to an unseen building environment. Although important to note, any DQN with a constant step-size (satisfying Robbins Monroe conditions [41]) towards the target is effective in nonstationary environments as it gives more weight to recent rewards [42]. They consider a building with multiple Variable Air Volume (VAV) zones. They have two test cases: a single RL agent controlling a single zone (while the rest operate a fixed schedule), and five RL agents operating simultaneously. Their research is also developed assuming that these non-stationarity events last **an entire week**, making it difficult to assess robustness on realistic outages, occupant behavior, extreme weather, etc. The convergence takes around 100 iterations to complete.

Fang et al. [28] introduces a single DQN that controls the supply air temperature and chilled supply water temperature set points of a VAV system. They decide to define the action space as follows: seven discrete temperatures for both supply air temperature and supply water temperature (7 x 7 = 49 possible actions). A thorough evaluation was done on only one month of training (July) and one month of testing (August). They conclude their study with a call to extend research on higher action spaces, and to deploy the controller in the real built environment.

Wang et al. [5] propose a similar approach to [28]. Here, they implemented a lightweight Deep Double Q-Network (DDQN) [43] for multi-zone, open office units. This environment is similar to the one from X. Deng et. al [35]. They also introduce a signal smoothness control term to prevent overuse of the system from frequent on/off transitions. The reward is a simple function of energy usage, temperature comfort violation, and the signal smoothness term. This approach performs well on a building model, simulated across weather files sourced from only inside the United States (Texas, California, South Carolina, etc.).

1.3 Background

The background section provides an overview of the foundational concepts and methodologies essential to understanding the optimisation of HVAC systems through RL. It introduces key notations and defines the Markov Decision Process (MDP) framework, which is central to RL. The discussion extends to specific RL algorithms, such as DDQNs, highlighting their relevance to controlling complex environments like multi-zone HVAC systems. Additionally, the section touches on the importance of simulation environments, such as EnergyPlus, which play a critical role in training RL agents by modeling building dynamics and energy consumption.

1.3.1 Notation

As a field, RL relies on an understanding of the relevant notation (Table 1.1). We will introduce the basics of RL but limit further discussion to only the relevant algorithm, DDQN, under which the HVAC control agent in this study operates.

Symbol	Meaning
$\mathcal{P}$	State Transition Probability Matrix
$\mathcal{H}$	History
s, s'	States
a	Action
r	Reward
$\mathcal{S}$	Set of all nonterminal states
$\mathcal{S}^+$	Set of all states, including the terminal state
$\mathcal{A}(s)$	Set of all actions available in state s
$\mathcal{R}$	Set of all possible rewards, a finite subset of $\mathbb{R}$
γ	Discount Factor $\gamma \in [0, 1]$
α	Step Size $\alpha \in [0, 1]$
$\mathbb{E}[X]$	Expectation of a random variable X
$\in$	Element of; e.g., $(s \in \mathcal{S}, r \in \mathcal{R})$
t	Discrete time step
A_t	Action at time t
S_t	State at time t, typically due, stochastically, to S_{t-1} and A_{t-1}
R_t	Reward at time t, typically due, stochastically, to S_{t-1} and A_{t-1}
G_t	Return following time t
π	Policy (decision-making rule)
$\pi(s)$	Action taken in state s under deterministic policy π
$\pi(a\|s)$	Probability of taking action a in state s under stochastic policy π
$p(s', r\|s, a)$	Probability of transition to s' with reward r from s and action a
$p(s'\|s, a)$	Probability of transition to s' from state s taking action a
$r(s, a)$	Expected immediate reward from state s after action a
$v_\pi(s)$	Value of state s under policy π (expected return)
$v_*(s)$	Value of state s under the optimal policy
$q_\pi(s, a)$	Value of taking action a in state s under policy π
$q_*(s, a)$	Value of taking action a in state s under the optimal policy
θ, θ_t	Parameter vector of value function

Table 1.1: Summary of Notation

1.3.2 Markov Decision Process (MDP)

Reinforcement learning is best understood as a Markov Decision Process, or MDP. It is a useful mathematical framework to present sequential decision-making problems. This process is often described similarly to Figure 1.4 [42], with states appearing as circles, certain actions appearing as dots, and the rewards for taking the action annotated nearby. Probabilistic actions will also contain the corresponding annotations next to the possible transition paths.

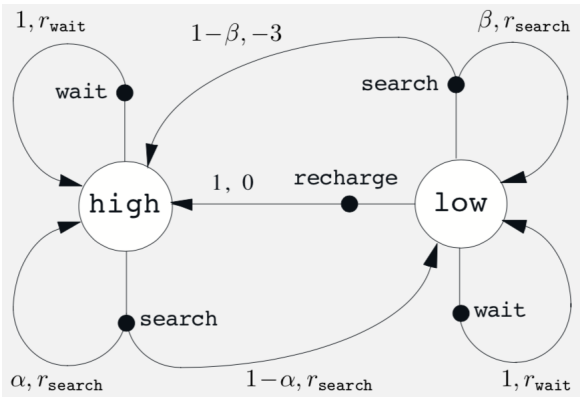


Figure 1.4: Example Markov Decision Process

The MDP is a foundational concept in reinforcement learning, predicated on the Markov property which asserts that "the future is independent of the past, given the present." This principle is formally expressed as in equation 1.1.

$$p(r, s \mid S_t, A_t) = p(r, s \mid \mathcal{H}_t, A_t) \tag{1.1}$$

This allows us to consider an agent's state as containing all of the useful information from the history and adding more history will not improve our situation.

The MDP is a tuple: $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$. It relies on transition probabilities and a reward function.

$$P^a_{ss'} = P[S_{t+1} = s' \mid S_t = s, A_t = a] \tag{1.2}$$

$$R^a_s = \mathbb{E}[R_{t+1} \mid S_t = s, A_t = a] \tag{1.3}$$

The transition probability is defined by equation 1.2, and the reward function is given by equation 1.3.

1.3.3 Reinforcement Learning (RL)

The central goal in reinforcement learning is for an agent(s) to optimise for a sum of rewards through repeated interaction with an environment - in our case, a building. The sum of these rewards is known as a return. RL is a flexible framework as we can consider any goal as an outcome of maximising a cumulative return [42].

The four foundational pieces of RL are the policy, value function, reward signal, and (optionally) the model of the environment. This agent behavior is formalised through a policy (equation 1.4), or distribution over actions for a given state s.

$$\pi(a|s) = P[A_t = a \mid S_t = s] \tag{1.4}$$

We can also define the value of a state as the sum of expected rewards the agent will collect as it moves from the current state further into time. The rewards accumulated from that sequence are known as a return (equation 1.5).

$$G_t = R_{t+1} + R_{t+2} + R_{t+3} + \cdots + \sum_{k=0}^{\infty} R_{t+k+1} \tag{1.5}$$

In certain applications (e.g., finance) it is useful to consider earlier rewards as more

valuable. Therefore, it is common to define the return as a sum of discounted rewards ("Myopic" if $\gamma = 1$, no discounting if $\gamma = 0$). This is formalised in equation 1.6.

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots + \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \tag{1.6}$$

We can now describe the value of a state with the following state-value functions, seen in equation 1.7. Formally, the state-value function is defined as the expected return from state s, following a policy π.

$$v_\pi(s) = E_\pi[G_t \mid S_t = s] \tag{1.7}$$

We can also extend this to consider the value of being in a state and taking a corresponding action, seen in equation 1.8. This state-action-value function (Q-function) is defined as the expected return from state s, taking action a and following a policy π.

$$q_\pi(s, a) = E_\pi[G_t \mid S_t = s, A_t = a] \tag{1.8}$$

Expanding the expectation we can view the state-action-value function more intuitively, as shown in equation 1.9.

$$q_\pi(s, a) = r(s, a) + \sum_{s'} p(s'|s, a) \left[\gamma \sum_{a'} \pi(a'|s') G_{t+1}(s') \right] \tag{1.9}$$

Richard Bellman, the father of dynamic programming (DP), created a recursive decomposition of both value functions. This decomposition equates the value of a state under a given policy to the value of the subsequent states. This concept, basing a value estimation off of another estimation is commonly known as bootstrapping. Thus we can consider the following Bellman Expectation equation [44] as bootstrapping off of the current value

estimate of the next state as seen in equation 1.10.

$$q_\pi(s,a) = r(s,a) + \sum_{s'} p(s' \mid s,a) \left[\gamma \sum_{a'} \pi(a' \mid s') q_\pi(s',a') \right] \tag{1.10}$$

Since we can have a large amount of policies for a single MDP, we need to know when we have found the optimal one. As stated earlier, the goal of reinforcement learning is to optimise for a sum of rewards through repeated interaction with the environment. We can consider the optimal action-value function (equation 1.11), over all possible policies, as the solution to that learning problem.

$$q_*(s,a) = \max_\pi q_\pi(s,a) \tag{1.11}$$

As seen in the Bellman Optimality Equation, equation 1.12, this solution is difficult and sometimes impossible to compute directly. Therefore an iterative update towards a solution is valuable.

$$q_*(s,a) = r(s,a) + \sum_{s'} p(s' \mid s,a) \left[\gamma \max_{a'} q_*(s',a') \right] \tag{1.12}$$

We can use dynamic programming to approximate a solution. We can turn equation 1.12 into an update, building an algorithm for control known as value iteration (equation 1.13).

$$q_{k+1}(s,a) \leftarrow r(s,a) + \sum_{s'} p(s'|s,a) \gamma \max_{a'} q_k(s',a') \tag{1.13}$$

The updates will reach a natural end when $q_{k+1}(s,a) = q_k(s,a)$, as we will have found the optimal value function q_* (and by extension, the optimal policy).

This is a model-free algorithm as we do not rely on knowledge of the underlying MDP. However, we can estimate it through sampling, as seen in temporal difference (TD)

learning [45]. We derive the Q-learning update by Bootstrapping on our estimate of the value of the state, compared to what we observe [46]:

$$q_{t+1}(s,a) \leftarrow q_t(S_t, A_t) + \alpha_t \left[r(S_{t+1}, A_{t+1}) + \gamma \max_{a'} q_t(S_{t+1}, a') - q_t(S_t, A_t) \right] \quad (1.14)$$

Standard Q-learning has a positive bias as the same Q-value function is used to estimate and select actions taken. A simple but powerful solution to this is to split into two Q-networks. One Q-network is tasked with providing the action the behavior policy takes (Q_A), and the other is tasked with determining the optimal action to take (known as the target). This is known as Double Q-learning [47], shown in Algorithm 1. Note: the target network is periodically synced with the action network.

Algorithm 1 Double Q-learning

Initialise: Q^A, Q^B, s
repeat
  Choose a, based on $Q^A(s,\cdot)$ and $Q^B(s,\cdot)$, observe r, s'
  Choose (e.g. random) either UPDATE(A) or UPDATE(B)
  if UPDATE(A) **then**
    Define $a^* = \arg\max_a Q^A(s', a)$
    $Q^A(s,a) \leftarrow Q^A(s,a) + \alpha(s,a)\left(r + \gamma Q^B(s', a^*) - Q^A(s,a)\right)$
  else
    Define $b^* = \arg\max_a Q^B(s', a)$
    $Q^B(s,a) \leftarrow Q^B(s,a) + \alpha(s,a)\left(r + \gamma Q^A(s', b^*) - Q^B(s,a)\right)$
  end if
  $s \leftarrow s'$
until end

This is the base algorithm applied to our HVAC energy management problem. However, the tabular case will be unable to scale to the size of our application as there will be too many states (rooms, zones, temperature ranges, etc.) to store in memory. It will be impractical to learn the values of each state separately.

We will therefore generalise across states with tactics from function approximation using Q-Networks. Specifically deep neural networks approximate the Q-functions parameterised with θ. A Q-Network can be trained by minimising a sequence of loss functions $L_i(\theta_i)$ that change at each iteration, i, shown in equation 1.15.

$$L_i(\theta_i) = \mathbb{E}_{(s,a)\sim\rho(\cdot)}[(y_i - q(s,a;\theta_i))^2] \tag{1.15}$$

Here, y_i is the target at i, and $\rho(\cdot)$ is the behavior distribution (equation 1.16).

$$y_i = \mathbb{E}_{s'\sim\epsilon}[r(s,a) + \gamma \max_{a'} q(s',a';\theta_{i-1}) \mid S_t = s, A_t = a] \tag{1.16}$$

It is important to mention the weights are fixed at the start of each iteration, but they adapt to form the target (as opposed to traditional supervised learning). We can optimise by taking the gradient of the loss (equation 1.17). This calculation allows us to update our parameters in the direction opposite of the loss using techniques like stochastic gradient descent (SGD):

$$\nabla_{\theta_i} L_i(\theta_i) = \mathbb{E}_{(s,a)\sim\rho(\cdot);s'\sim\epsilon}\left[(r + \gamma \max_{a'} q(s',a';\theta_{i-1}) - q(s,a;\theta_i))\nabla_{\theta_i} Q(s,a;\theta_i)\right] \tag{1.17}$$

This is known as Deep Reinforcement Learning, and we have arrived at the foundational algorithm chosen in this paper, the DDQN [43].

DDQNs also include an experience replay module, allowing the RL agent to be simultaneously model-based, with sampled transitions and rewards as a non-parametric model, and model-free. The experience replay module also allows us to adhere to the traditional deep learning assumption of samples from our training data being independent and identically distributed (IID). This is because we sample from random batches of transitions, drawn from the same distribution.

Now that we have established the background material to understand the agent, we will introduce the simulation and real training environments as well as the dynamics of building energy management.

The value of a simulator or other modeling tactics is predicated on the DRL training environment, specifically for HVAC management. Training a DRL agent to converge in these complex environments requires modeling countless transitions using thermodynamics and HVAC component interactions, all while characterising their corresponding consumption. This makes most applications without simulation environments, such as EnergyPlus, challenging.

1.3.4 EnergyPlus

EnergyPlus is a whole building energy simulation program from the United States Department of Energy (DOE) that we use to model energy consumption. It is capable of modeling energy from heating, cooling, ventilation, lighting and plug and process loads as well as water use in building models [15].

EnergyPlus simulations use input data files (IDF) to pass all the information about the building model (geometry, materials, schedules, HVAC systems, etc.). The IDF file must reference an EnergyPlus weather data file (EPW).

After setting simulation parameters (start/end date, time-step, etc.) the simulation is run to completion.

1.3.5 HVAC operation

HVAC systems are designed to provide control over the indoor environment while ensuring occupancy comfort. The principal component in this system is the Air Handling Unit (AHU). The AHU intakes outside air, filters it to remove dust or other particles, and heats or cools it depending on the status of the BEMS. In zone-based HVAC systems, outside

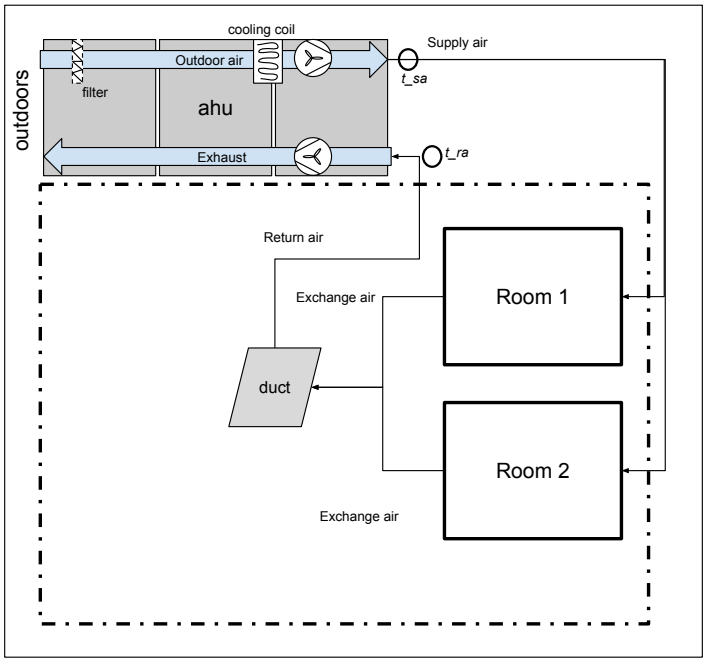


Figure 1.5: HVAC System High-Level Diagram - AHU

air is conditioned and distributed throughout the building via ducts. VAV units control the volume of air flowing into their corresponding zones and adjust the temperature using reheat coils. This ensures that each area maintains the desired temperature and meets the building's quality standards. VAV systems typically achieve better heat transfer efficiency because they can precisely regulate airflow and temperature in response to the specific needs of each zone, thereby optimising energy use and enhancing comfort [48]. A simplified, system-level diagram is shown in Figure 1.5.

2 | Methodology

This chapter introduces the components used in the experiments, including models, metrics, and datasets. Our experiments utilise modified versions of Double Q-Networks for multi-zone HVAC optimisation across various climates. We apply a novel replay buffer mechanism, Replay across Experiments (RaE) [4], and evaluate our models using standard performance metrics. These models are applied to multi-zone HVAC optimisation across diverse climatic conditions using a real-world dataset of global weather data from Singapore, UAE, UK, US, Canada, etc. Performance is evaluated by measuring energy consumption and comfort violation.

2.1 Models

The models used in our experiment are modified versions of the Deep Double Q-Network (DDQN) presented in Wang et al. (2023) [5] . The DDQN mitigates the overestimation bias in Q-learning by maintaining two separate value functions. One network is used to select actions, and the other to evaluate them. This mechanism, along with the replay buffer, helps stabilise the training process.

The action space is binary for each zone, with 1 enforcing temperature control, effectively turning the HVAC systems "ON", and 0 mapping to the systems being out of use, or "OFF".

We first attempted to reproduce the DDQN specifically tailored for open offices [5]. We evaluated the baselines for four climates (AZ, TX, SC, and CA) and measured the performance. We found it similar to the results reported for energy savings, however, we were unable to replicate the low comfort violation percentages reported. To remedy this, we made some necessary changes to the algorithm they provided. We refactored and added the exploratory parameter, epsilon, which was previously not included.

We trained this DDQN in each climate for 20 epochs with the hyperparameters seen in Table A.1. This training process took approximately 120 minutes per DDQN. We then compared energy savings results to the previous baseline reported values as seen in Table 3.1.

Parameter	**Value**
timestep_per_hour	12
state_dim	15
action_dim	64
epochs	20
lr	0.001
gamma	0.9
epsilon	0.003
epsilon_min	0.001
epsilon_decay	0.95
target_update	200
buffer_size	10000
minimal_size	200
batch_size	128
T_factor_day	0.02
E_factor_day	1e-06
T_factor_night	0
E_factor_night	1e-06

Table 2.1: DDQN Training Parameters

These pre-trained DDQNs are used in all further experiments as a starting point. We load pre-trained model weights for both the target and actor Q-network at the start of the evaluation epoch. From this point, we accelerate our evaluation by measuring performance on a single epoch.

We extend these DDQNs to include a growing dataset that fuses with the replay buffer. This dataset contains context from prior experiments, described in Replay Across Experiments, or, RaE [4]. As we train in different climates, our agent has access to experience from previous experiments. This is mixed into the replay buffer of the current experiment

at a fixed ratio (1:1).

We present two sizes of prior datasets: 10k and 100k sampled transitions. Note: at 12 time steps an hour, 100k transitions cover nearly an entire epoch of training. The differences in model architectures are summarised in Table 2.2.

Category	base DDQN	RaE DDQN	RaE DDQN (large)
Parameters	607,360	607,360	607,360
Layers	3	3	3
Hidden Size	512	512	512
State Dimension	15	15	15
Action Dimension	64	64	64
Replay Buffer Size	10,000	10,000	10,000
RaE Dataset Size	n/a	10,000	**100,000**

Table 2.2: Model Architecture and Performance Benchmarks

We begin measuring performance by defining the internal function for our scalar reward signal. This function is composed of three components (energy, comfort, signal loss), each helping us quantify performance in higher-level tasks.

2.2 Reward function

The reward function provides the agent with a scalar measure of success, according to the actions taken in the environment. We use the reward function defined by Wang, et. al [5] as a negative sum of total losses; therefore, the agent's updates should move towards minimising the loss function defined in equation 2.1. This achieves both energy optimisation and thermal comfort. The reward function also considers a few constraints to accommodate real-world requirements, such as setting reasonable upper and lower bounds for HVAC set points to prevent health or safety issues and incorporating a term to penalise frequent ON/OFF transitions to reduce the mechanical wear of the ventilation system.

$$R = -L_{\text{total}} = -(L_{T_i} + L_E + L_S) \tag{2.1}$$

The first term, L_{T_i} represents the temperature loss, defined in equation 2.2.

$$L_{T_i} = \begin{cases} 0 & \text{if } T_{\min} \leq T_i \leq T_{\max} \\ \eta_T \cdot (T_i - T_{\min})^2 & \text{if } T_i \leq T_{\min} \\ \eta_T \cdot (T_i - T_{\max})^2 & \text{if } T_i \geq T_{\max} \end{cases} \tag{2.2}$$

Here, T_i is the temperature of zone i, $T_{\min}$ and $T_{\max}$ are the minimum and maximum desired temperatures, and η_T is a tunable temperature loss factor. The values of which are taken from Wang, et. al [5]. This formulation ensures a zero reward if the current temperature is within the desired range and a negative reward proportional to the distance to the center of the desirable range when the temperature is outside the range.

The second term, L_E, is the energy consumption loss, defined in equation 2.3. Here, E_t is the energy consumption at time t and η_E is the energy loss factor. This term is post-normalised to account for the different primary scales of energy loss and comfort loss.

$$L_E = \eta_E \cdot E_t \tag{2.3}$$

The final term, L_S, is the smoothness loss, defined in equation 2.4. Here, $A_{i,t}$ is the action of VAV unit i at time step t, $\oplus$ denotes the XOR operation, and η_S is the smoothness loss factor. This term penalises frequent ON/OFF transitions to reduce mechanical wear.

$$L_S = \eta_S \cdot \sum_{i=1}^{n} (A_{i,t} \oplus A_{i,t-1}) \tag{2.4}$$

Note: although a monotonically increasing cumulative return (sum of discounted re-

wards) typically indicates convergence in reinforcement learning, the dataset's changing seasons cause difficulties towards the end of the year, even as the agent improves its value approximations. This is not accounted for in the training process or the reward function.

2.3 Datasets

2.3.1 Building model

The building model, "ITRC 2nd 6-zone OPEN," designed as a multi-VAV digital twin [5], remained common across our experiments. It contains six zones, controllable through VAV units.

2.3.2 Weather

Our experiments use real-world weather data from seven countries: England, Canada, South Africa, Japan, the United Arab Emirates, Singapore, and the United States. These datasets provide a comprehensive range of climate conditions, allowing us to test the generalisability and robustness of our models. Each dataset includes historical weather data spanning a full calendar year, with each month sampled from a different year.

An overview of the key characteristics of the datasets is depicted in Appendix Table A.2.

The main task of this thesis is to evaluate the performance of the DRL agent in each of these environments. Therefore we chose weather patterns sourced from mainly extreme climates (with the exception of California). They span a large range of temperatures as shown in Figure 2.1. The temperature displayed is the dry-bulb temperature (DBT), or the air temperature as measured by a thermometer not directly exposed to radiation. The relative humidity (RH) is plotted on the same plot expressed as a percentage. It is a measure of the actual amount of water vapor in the air compared to the total amount of vapor that can possibly exist in the air at current temperatures.

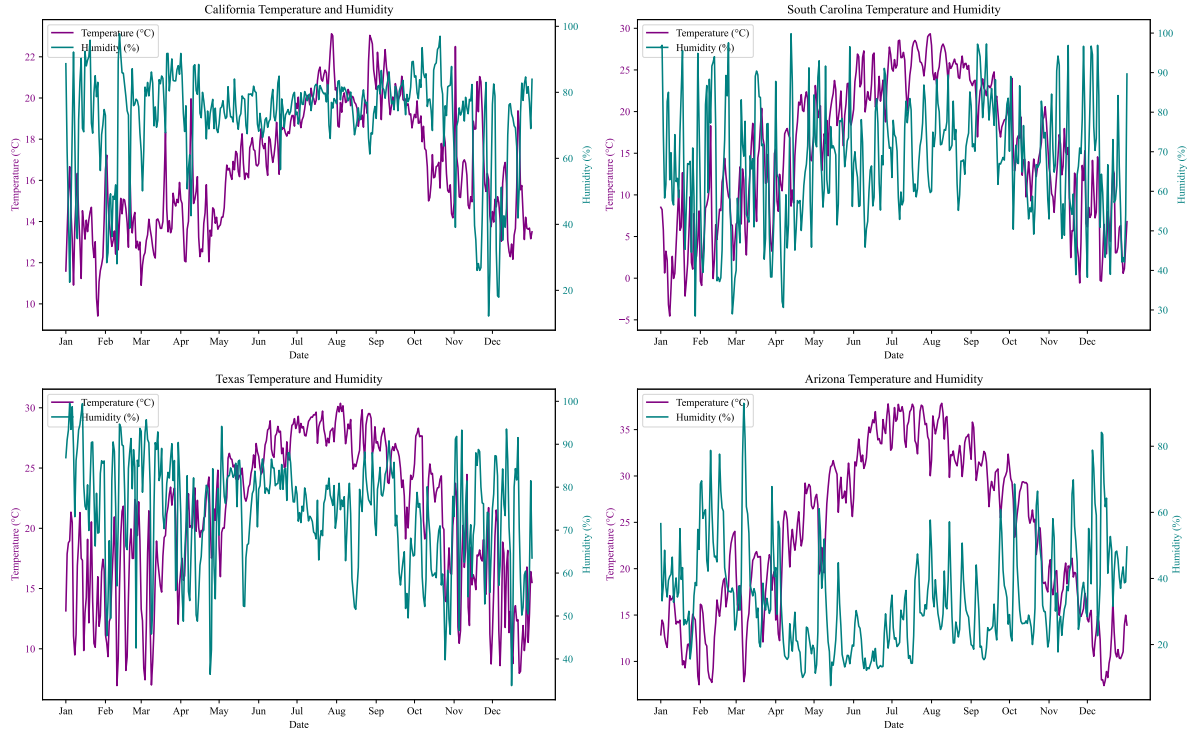


Figure 2.1: Exploratory Data Analysis - US Weather Files (DBT and RH)

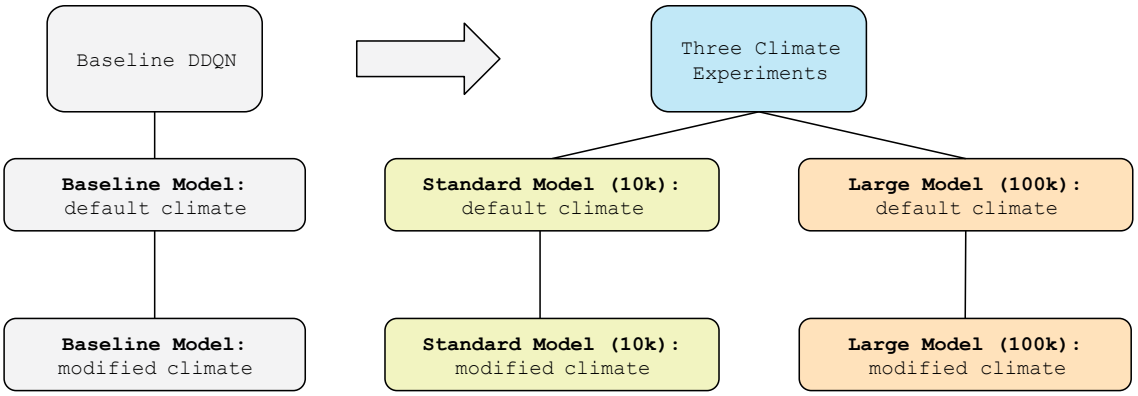


Figure 2.2: Overview of Experiments

At present, the weather files do not represent the current temperatures experienced. For example, in Texas, much of May through August has seen temperatures above 32 °C (Appendix Figure A.1), whereas the climate files average much lower temperatures during that period (Appendix Figure A.2). This is addressed in the second half of the Three Climate Experiments.

2.4 Experiments

In the following section, we discuss the relevant conditions of our experiments as well as the tasks and metrics that we use to quantify performance. The overview of our experiments is summarised in Figure 2.2.

2.4.1 The Three Climate Experiments

Our first task is measuring the relationship between prior knowledge from a growing amount of experience on the speed of convergence to optimal strategies for reducing energy consumption. Here, we define an experiment as a training run of one calendar year while in the same climate.

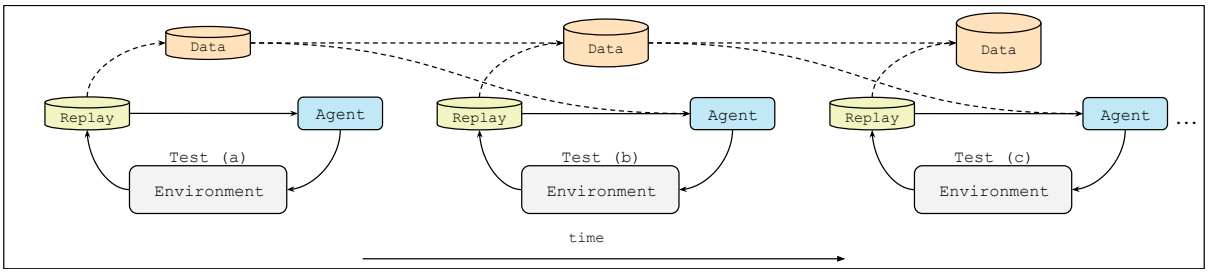


Figure 2.3: Three Climate Experiment Framework

We execute experiments in the sequence "Test A", "Test B", and "Test C" as seen in Figure 2.3. Samples from the replay buffer of the current experiment are carried over from test to test. In the standard RaE model, 10,000 samples are carried over from Test A to Test B and used to update the agent in the current experiment. Then the same 10,000 samples are combined with the replay buffer generated in Test B to form a database of 20,000 samples. This database is then used in Test C to update the agent. In the large RaE model behavior is the same, except 100,000 samples are carried over, and Test C is run with a database containing 200,000 transitions. This behavior and experiment flow is described in Algorithm 2.

Algorithm 2 Replay across Experiments

```
Initialise Variables: climate list, experiment length, epochs, buffer size, sample ratio
for each climate in climates do
    Initialise a primary replay buffer for the climate
    for each experiment index up to the length of experiments for that climate do
        Initialise the agent
        if experiment index is 0 then
            Run the experiment no secondary replay buffer
        else
            Initialise a secondary replay buffer
            Load samples from all previous experiments into the secondary buffer
            Run the experiment, sampling from both buffers at sample ratio
        end if
        Save the results of the current experiment
        Save the replay buffer of the current experiment
    end for
end for
```

These experiments were run on the base DDQN, RaE DDQN, and RaE DDQN (Large) models. To baseline these results from RaE we first ran "random weight resetting" experiments [49] and measured energy savings and comfort violation. To isolate the effect of weight resets from the benefits of data reloading (as in RaE), we consider a baseline where all weights (target and actor) are reset every K policy update. Here, K is set to the number of updates used to train the policy that generated the data. We experiment with reset frequencies of K, K/10, and K/100. All experiments are run until convergence or at least 2 x K updates.

We then manipulate default weather data files to better represent the present climate. This is achieved by adding Gaussian noise to the dataset to introduce variability that mirrors natural fluctuations. Specifically, the noise is parameterised to reflect realistic deviations observed in historical weather data. Additionally, we adjust the extreme values within the dataset by 10%, to simulate the intensified weather events occurring at present.

Next, we apply these modifications to dry bulb temperature, humidity, wind speed, and solar radiation, ensuring that the alterations are consistent with both empirical data and projected trends. By carefully calibrating the level of noise and the degree of extreme manipulation, we aim to create a data set that more accurately reflects the complex interplay of variables in a changing climate.

Furthermore, the manipulated data undergoes a validation process where the generated dataset is compared against recent observational data to ensure the modifications result in realistic representations (Appendix Figure A.3). This comparison is crucial to show that the simulated present climate is not only statistically plausible but also consistent with the observed trends. As a result of this comparison, we trimmed the RH values towards only physically possible values (RH <100%).

2.4.2 Metrics

TD-loss

In the DQN training process, we are attempting to minimise the TD loss to converge towards a stable model of our Q-value functions. This is formally defined in equations 1.15 and 1.16. TD-loss (equation 2.5) is defined as the Mean Squared Error (MSE) difference between the predicted Q-values and the actual Q-values, where N is the number of samples, $Q_{\text{target},i}$ is the target Q-value for the i-th sample, and $Q_{\text{predicted},i}$ is the predicted Q-value for the i-th sample.

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(Q_{\text{target},i} - Q_{\text{predicted},i})^2 \tag{2.5}$$

Energy Consumption (EC)

The first building-specific metric we are considering is energy consumption. This is the sum of energy consumed by the HVAC systems across the simulation run shown in equation 2.6. This is compared between RBC and DRL-based strategies in equation 2.7, where EC is the total energy consumption, P_t is the power consumption at time t, and Δt is the time interval.

$$EC = \sum_{t=1}^{T} P_t \Delta t \tag{2.6}$$

Energy Savings (ES)

$$\text{Energy savings \% (HVAC)} = \frac{EC_{\text{RBC, HVAC}} - EC_{\text{DRL, HVAC}}}{EC_{\text{RBC, HVAC}}} \times 100 \tag{2.7}$$

Energy Savings (ES) is defined in equation 2.7, where $EC_{\text{RBC, HVAC}}$ and $EC_{\text{DRL, HVAC}}$ are the energy consumptions of the HVAC system under RBC and DRL-based strategies respectively.

Comfort Violation (CV)

The comfort violation is calculated after an epoch is finished as shown in equation 2.8. It is a measure of what percentage of time an occupant would be likely to experience discomfort.

$$CCR = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} I(T_i^{\min} \leq T_{i,t} \leq T_i^{\max}), \tag{2.8}$$

Here, CCR is the comfort compliance ratio, T_t is the temperature at time t, $T_{\min}$ and $T_{\max}$ are the minimum and maximum desired comfort temperatures, respectively, and

$\mathbf{1}\{\cdot\}$ is an indicator function. For all the experiments in this study $T_{\text{min}} = 21.6°\text{C}$ and $T_{\text{max}} = 23.3°\text{C}$.

We report this value as a percentage of time in discomfort, as seen in equation 2.9.

$$\text{Comfort Violation} = (1 - CCR) \times 100\% \tag{2.9}$$

3 | Results and Analysis

In this chapter, we summarise the experiments and results of our project. We demonstrate the baseline performance of the DDQN. Then, we test the DDQN's performance across experiments, showing a positive correlation between the amount of experience and performance across the evaluation period as validated on multiple dataset sizes. Finally, we evaluate the generalisation capability and robustness of the proposed DDQN model with RaE by testing on modified climates, outside of the training distribution of the model.

3.1 Model evaluation

The DDQN models were evaluated on the HVAC simulation data using the building scenario from [5]. Computational resources included a 2023 MacBook Pro with an M3Pro Chip and 18GB of memory, utilising an MPS backend to accelerate training. The models were assessed based on standard RL metrics such as average reward and MSE loss over time. The models were further assessed using energy consumption and thermal comfort metrics.

3.1.1 Performance metrics

We first attempted to reproduce a DDQN [5] to establish a baseline. We evaluated the model with standard performance metrics for Arizona, Texas, South Carolina, and California. We found it similar to the results reported for energy savings, however, we were unable to replicate the low comfort violation percentages reported. We then made some necessary changes to the algorithm provided by Wang et al. [5], refactoring the code to an object-oriented, modular approach. We added a random seed for statistical significance and reproducibility. We also made the code for the project available as fully

open-sourced. Also, the exploratory parameter epsilon was added, along with minor bug fixes. The energy savings results, compared to their reported values are shown in Table 3.1. Due to the inherent trade-off between energy consumption and thermal comfort, the metrics are reported separately.

	AZ	**TX**	**SC**	**CA**
ES (reported) (%)	27.82 / 28.77	24.14 / 25.56	33.96 / 38.83	38.56 / 47.36
CV (%)	5.74 / 2.55	5.30 / 2.40	3.47 / 0.51	3.41 / 0.36
Refactored ES (%)	23.63 / 28.77	22.64 / 25.56	28.11 / 38.83	40.39 / 47.36
Refactored CV (%)	2.49 / 2.55	2.64 / 2.40	0.75 / 0.51	0.71 / 0.36

Table 3.1: Energy Savings and Comfort Violation Compared to Previously Reported Results

We used TD-error, converted to MSE loss, to measure DDQN convergence. The model converges quickly after only around 50,000 sampled transitions (6 months). This stability is then tested towards the end of the year as the weather in most climates becomes more extreme. This can be mitigated by training over 20 epochs.

The global weather files were not previously reported. We report standard performance metrics on this novel data in Table 3.2.

	Cape Town	**London**	**Dubai**	**Tokyo**	**Vancouver**
ES (%)	40.77	36.59	10.68	28.86	42.76
CV (%)	0.61	0.59	8.19	1.63	1.10

Table 3.2: Energy Savings (ES) and Comfort Violation (CV) Metrics for Global Weather Locations.

3.2 Replay across Experiences (RaE) evaluation

3.2.1 Baseline

We first baseline RaE performance by running random weight resetting [49] experiments and measuring ES and CV (Table 3.3).

We reset the weights of the second and third layers with Kaiming uniform initialisation specifically designed for layers with ReLU activation [50]. We chose this method to keep the scale of the gradients roughly the same in all layers. It helps maintain the variance of the activations throughout the layers. It samples weights from a uniform distribution bounded by $[-\sqrt{\frac{6}{\text{fan\_in}}}, \sqrt{\frac{6}{\text{fan\_in}}}]$, where "fan_in" is the number of input units in the weight tensor.

We measure the impact of random weights by reinitialising every "K" times, where K is set to the number of updates used to train the policy that generated the data. We therefore tested K, K/10, and K/100 cases with the experiments run to convergence.

Suppose we consider the first re-initialisation, at K frequency. This occurs in the middle of the evaluation of a pre-trained DDQN. As seen in Appendix Figure A.4 it only takes 40 timesteps to re-stabilise, corresponding to two updates of the target Q-network. The total results from the experiments are reported in Table A.3.

	K	**K/10**	**K/100**
ES (%)	17.89 [16.19, 19.58]	17.39 [15.74, 19.04]	12.70 [9.75, 15.64]
CV (%)	8.19 [7.46, 8.93]	8.11 [7.57, 8.64]	8.55 [8.21, 8.89]
Average Reward	-3.06 [-5.23, -0.89]	-5.09 [-5.13, -5.05]	-5.39 [-5.47, -5.32]

Table 3.3: Energy Savings, Comfort Violation, and Average Reward for K, K/10, and K/100 (with 95% confidence intervals)

3.2.2 The Three Climate Experiment

The results of our Three Climate Experiment demonstrate the effectiveness of our proposed approach in improving HVAC energy efficiency across diverse climatic conditions. A full table of results, including energy savings, comfort violations, and improvements over the baseline, is provided in Appendix Table A.4.

We present a snapshot of the ES, CV, and percentage point improvements over baseline for both default and modified climate parameters when applied in Vancouver, BC, in

Table 3.4. The Three Climate Experiment outperforms the base DDQN by **18%** in the more extreme environment, and by **54.11%** when introducing the larger dataset of 100k transitions.

Model Type	Climate	ES (%)	CV (%)	Δ ES Improvement
RaE	default	39.86	0.65	-5.37
RaE	modified	32.89	0.89	**18.13**
RaE (Large)	default	70.17	1.83	24.94
RaE (Large)	modified	68.88	1.61	**54.11**

Table 3.4: Three Climate Experiment Snapshot: Effect of Climate Modification on Energy Savings (ES), Comfort Violation (CV), and Improvement Over Baseline - Vancouver, BC

A more detailed view of our control schedule shows the differences introduced by the additional dataset at experiment runtime. We see that for a sample day (Figure 3.1) the addition of the experimental prior data helps to **smooth** the control signals for the HVAC settings in these zones as well as help to identify the **centroid zone** (3) where the temperature impact per kWh expended is higher than others. We can baseline the additional dataset introduced in RaE with random weight resetting, where we see significant variations in CV performance as the resetting frequency increases (Appendix Table A.3). For instance, random weight resetting led to **3.2x** higher comfort violations than the Three Climate Experiment in Tokyo. While random weight resetting improved ES performance, none of the experiments outperformed the Three Climate Experiment (Large) model.

For the Three Climate Experiments tested in modified climates, several key observations emerged. Our approach demonstrated substantial energy savings across climates, with ES reaching up to **70.17%**, and up to **68.88%**. However, comfort violations increased considerably when tested in desert climates, emphasising the importance of further climate-specific optimisation strategies (or relaxing comfort ranges in these climates). Our method outperformed all baselines, particularly when leveraging large existing ex-

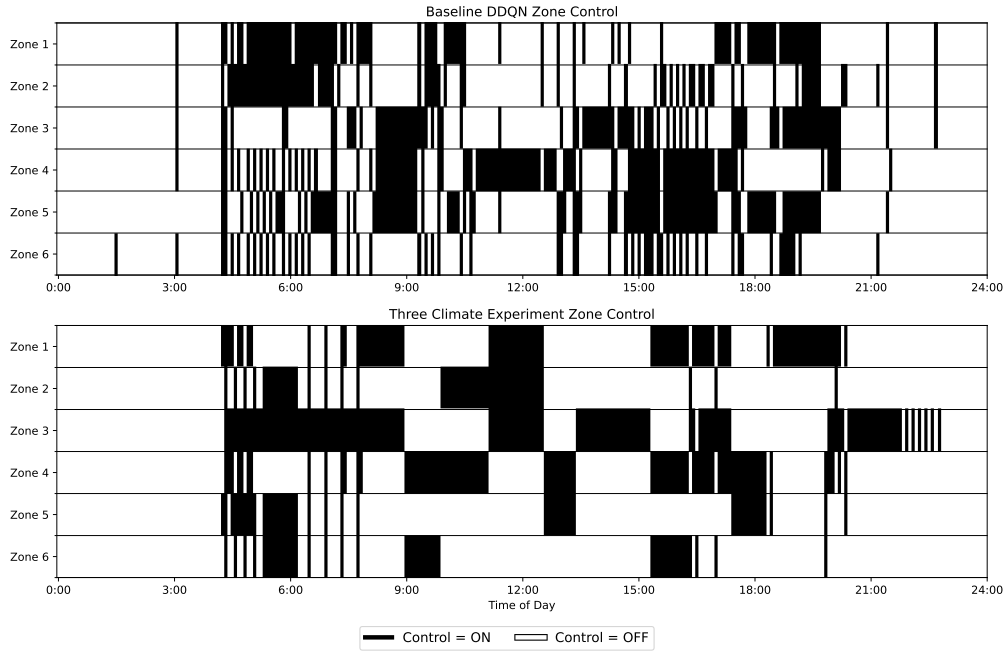


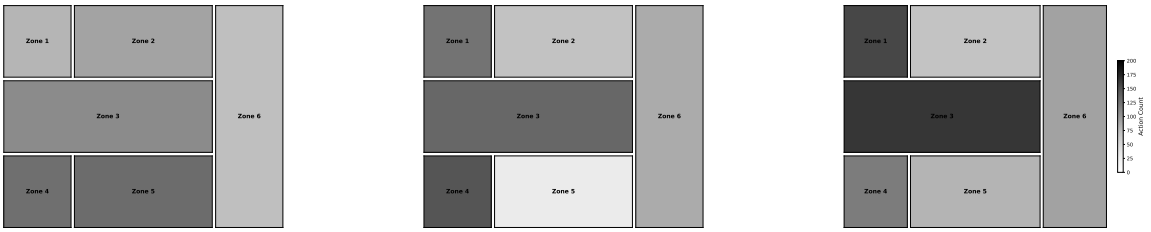
Figure 3.1: Six-Zone Control System: Example Day

periment datasets to move toward the optimal control solution. Notably, our approach led to significant improvements over the baseline DDQN method, achieving a **54.11%** increase in energy savings in Vancouver, BC (Table 3.4).

3.2.3 Explainability

As the results of this study are quite strong, with significant improvements in energy savings (over 70% as compared to the model predictive control baseline), a significant amount of work should be put into explaining these results and shaping them to be useful for real practitioners and building managers. It is not enough for the model to perform well, the decisions it makes must be understandable to those who will implement them in practice. For building managers, in particular, understanding how the model arrives at specific control actions is crucial for integrating these solutions into existing workflows and ensuring they align with operational goals.

In closed offices, heat gain in each zone is influenced by dedicated HVAC systems and heat conduction between spaces. In contrast, HVAC units in open offices affect not only their area but also neighboring zones. Thus, our solution's identification of a critical central zone, Zone 3, with potential ripple effects that enhance overall performance, is validated by Figure 3.2, particularly for Vancouver, which shows one of the best-performing trials.



(a) Baseline DDQN (b) Three Climate (c) Three Climate (Large)

Figure 3.2: Comparison of Action Distributions for Agents in Vancouver (modified climate). Darker Zones Represent More Counts of Enforcing Temperature Control.

Future work focusing on explainability could involve exhaustively analysing how the model's decisions vary across different climates and different model complexities (hidden dimension sizes). By understanding these variations, we could gain insights into why the model performs better or worse under certain conditions.

In summary, while the performance of the HVAC optimisation model is impressive, ensuring that the model's decisions are explainable is also important. A concerted effort was made to interpret and communicate these results effectively with real practitioners and building managers to utilise these advanced models confidently (Appendix Section B). This effort is aligned with working towards SDG 11 (Sustainable Cities and Communities).

4 | Discussion

In this discussion chapter, we critically examine the experimental findings and identify the limitations of our approach. Based on these limitations, we propose future work that could advance this research.

4.1 Findings and analysis

The Three Climate Experiment models outperform in varied environments and emerge as a resilient control solution capable of adapting to changing climate conditions. If we consider using 212,233 MJ per year as a baseline [5], we can model the impact of the energy savings observed in the study as a total of **1.064 million kWh p.a.**. This significant energy reduction demonstrates the potential of our approach in furthering SDG 7 (Affordable and Clean Energy).

Initially, we found difficulty reproducing baseline results. This was mainly due to the issues in the codebase and the fine-tuning required for hyperparameters that balance energy savings with comfort. After extensive experimentation, we successfully achieved similar energy savings in the climates reported in the original study (Arizona, Texas, South Carolina, and California). However, we encountered difficulties in replicating the low comfort violation percentages, suggesting that the balance between energy efficiency and occupant comfort remains a complex issue that variations may influence in hyperparameter settings and model architecture.

We expanded the evaluation to include climates that were not previously reported: Cape Town, London, Dubai, Tokyo, and Vancouver. Among these, only desert climates struggled significantly with maintaining occupancy comfort. Dubai, for example, achieved an **8.19%** comfort violation. This is a factor of eight difference compared to the other climates, most likely due to the extreme heat, suggesting indoor temperature comfort

ranges should be adjusted relative to these cases. These results highlight the varying effectiveness of the base DDQN approach across different climates, leading us to experiment with more adaptive and informed workflows.

The results of the Three Climate Experiment provide several insights into the effectiveness and challenges of applying DRL to HVAC optimisation across diverse climates. Contrary to our baseline results, we empirically found that RaE has very low sensitivity to hyperparameter choices across most diverse domains. Furthermore, the wide range of energy savings (5.67% to 69.52%) across climates emphasises the difficulty of this optimisation problem reaching beyond basic cross-validation. This variability reinforces the value of furthering our adaptive, learning-based systems towards a more contained solution. Therefore we sought to understand the real-world impact of our system. We engaged with building managers and industry professionals, providing us with valuable insights into the practical challenges of implementing such advanced control schedules into existing infrastructure, encouraging our approach towards smoother control schedules (Figure 3.1).

4.1.1 Dataset size

The impact of dataset size on HVAC optimisation performance is clearly demonstrated in the Three Climate Experiment. As shown in Appendix Tables A.4 and A.5, increasing the dataset size from 10k to 100k samples led to substantial improvements in energy savings across all tested locations. For example, in Vancouver, BC, energy savings increased from 39.86% to 70.17% under default climate conditions with the larger dataset, representing a 24.94% improvement over the baseline.

This improvement suggests that a larger dataset enables the model to capture more environmental variability and update in more informed ways. Sampling from an entire year of previous experiment data creates updates more robust to climate modifications. This

effect is consistent across climates, from Dubai's hot desert to London's temperate environment, underscoring the generalisability of the approach when supported by sufficient prior data.

While larger datasets significantly improve energy optimisation, they also exacerbate the challenge of maintaining occupant comfort, with the exception of extreme climates like Dubai (where the already high occupancy discomfort shows little variation). This is a significant limitation for the practical deployment of these models. Yet, there are positive examples such as London that achieve both high energy efficiency and low comfort violations under all test cases.

In the case of London, the energy savings under the 10k dataset in default climate conditions were 39.27%. Increasing the dataset to 100k samples improved energy savings to 69.52%, a notable 30.25 percentage point increase. Moreover, comfort violations remained low, decreasing slightly from 2.04% to 1.31%. In this case, the larger datasets significantly improved both energy efficiency and occupant comfort, demonstrating the potential of this approach in temperate climates where the trade-offs are less severe.

In contrast, in Dubai, the larger dataset led to a dramatic increase in energy savings from 10.09% to 55.41% while maintaining a similar comfort violation (8.139% as compared to 8.20%). Although there is a significant challenge of balancing energy savings and occupant comfort, these results suggest that larger datasets carried through experiments lead to improved efficiency and cause the RL agent to learn a policy capable of staying within a comfortable range.

4.1.2 Robustness to climate change

The Three Climate Experiment models demonstrate robustness to our modifications of the climate parameters, maintaining high energy efficiency with only minor reductions in performance across most locations. This is significant as the models have not experienced

these conditions in training.

As seen in Appendix Tables A.4 and A.5, the model maintained high energy savings even under altered environmental conditions, demonstrating its adaptability to changes in the climate. The robustness of the HVAC optimisation model was tested under modified climate conditions, comparing performance directly between default and modified climates using the same dataset size. This approach isolates the impact of climate variability, providing a clearer understanding of how well the model adapts to changing environmental conditions. Modifying the climate led to an average increase of **2.033** percentage points in comfort violation for standard datasets and **2.22** percentage points for large datasets.

Compared to the baselines, which saw much lower efficiency, the Three Climate Experiment models' ability to sustain high energy savings under both default and modified conditions highlights its effectiveness. For instance, in Vancouver, BC, with the 100k dataset, energy savings under default conditions were **70.17%**, which slightly decreased to 68.88% under modified conditions. This marginal reduction of **1.29** percentage points indicates that the model maintains high energy efficiency despite changes in climate. The baseline, however, saw a decrease of ES by **32%** as a result of the modification. Similarly, in California, the energy savings under default conditions were 69.67% compared to 68.45% under modified conditions, reflecting a minor decrease of 1.22 percentage points, further illustrating the model's stability when climate parameters are modified.

Tokyo provides another example where the model showed robustness with minimal changes in energy efficiency. With the 100k dataset, energy savings were **62.95%** under default conditions and 60.84% under modified conditions, a decrease of **2.11** percentage points. Although this decrease is slightly larger than in other temperate climates, it still indicates that the model can adapt well to changes in climate parameters without a substantial loss in efficiency. These comparisons underscore our models' robustness to climate modifications, particularly in temperate regions where the impact on energy efficiency is minimal.

Despite significant overall performance improvements, the findings highlight the need for explainability and further analysis to understand how the model can robustly handle extreme and variable climates.

4.1.3 Explainability

The explainability efforts in this study are a crucial step toward bridging the gap between our advanced models and their practical implementation in real-world settings. While our models have achieved significant energy savings, it is vital to translate these results into actionable insights that building managers can readily understand and apply. By examining the causality of our models' decisions, we have gained initial insights into its decision-making process. For instance, identifying Zone 3 as a critical area for HVAC control within specific climates underscores the importance of focusing on individual zones to drive overall building performance improvements compared to rule-based approaches.

Additionally, the **smoothness** of our solution suggests an improvement in control decisions, although this hypothesis requires further validation through exhaustive testing. Despite these promising efforts, explainability remains an area for continued exploration. Future research should focus on into understanding the model's behavior across diverse climates and configurations, with particular attention to why certain control decisions yield superior outcomes. Such insights would not only enhance the model's practical applicability but also build trust among practitioners, ensuring this solution can be confidently integrated into existing workflows.

4.2 General limitations and future work

This section discusses the broader limitations of our approach and provides directions that can help to address these limitations in the future.

Our experiments utilise only one core algorithm, the DDQN [47]. However, numerous

other models could enhance the performance of standard Double-Q Learning. Models focused on improving exploration and stability like Rainbow DQN [51] and Noisy DQN [52]. Models that are more data efficient like Prioritized DDQN [53]. Complete architectural modeling changes for more accurate value estimation like Dueling DDQN [54] and Distributional DQN [55]. Asynchronous parallel training in the case of A3C [56]. Testing these models could help us determine if we have the optimal solution and provide deeper insights into the underlying problem.

Additionally, our current research did not incorporate real-world industry factors such as energy prices, which are critical for practical applications. Future work could use these elements to make the solutions more applicable in real-world settings. Furthermore, we could extend our research to include a wider variety of building models to increase the generalisability and practical relevance of our findings. Our results suggest there is still a practical limitation in maintaining comfort in mixed environments so experimenting with applications such as Multi-Task RL could be a promising direction. We considered using Multi-Task RL which provides an index for this model to understand what climate it is in and perform either more dramatically or more conservatively. This could allow it to perform in a multitude of environments so each "task" could represent a different climate condition.

An example of the encoding is shown in Table 4.1. The extended state-space of the model would include this index and would learn a distinctly different control solution, potentially performing better at each task.

Task ID	**Climate**
1	Warm Marine
2	Mixed Humid
3	Hot Humid
4	Hot Dry

Table 4.1: Climate Task Mapping

5 | Conclusion

In this conclusion chapter, we summarise the research findings, address the main research question, and highlight the key contributions of this work. We also provide recommendations for future research and discuss the broader implications of this work in the field of HVAC optimisation and energy management.

5.1 Summary and reflection

The main research question addressed in this thesis was how to optimise HVAC control systems across diverse climates with DRL. The research aimed to efficiently manage energy in HVAC systems while maintaining occupant comfort, even under varying and extreme weather conditions. The answer to this question lies in the successful implementation of the RaE methodology combined with a modified DDQN in a framework that significantly improves the performance of HVAC systems across different climates.

This research presents a novel approach to HVAC optimisation that outperforms traditional RBC and current DRL methods. By incorporating the RaE methodology, the proposed system leverages experience from different climates, leading to improved energy efficiency and robustness. In the Three Climate Experiment, which trained the model across diverse climate conditions, we empirically found substantial improvements in energy savings (up to 54.11%).

This study addressed key limitations of existing DRL-based HVAC control systems, such as their inability to generalise across different climates and the high computational costs associated with their training. We successfully reproduced and improved a baseline DRL algorithm, leading to a modular, open-sourced framework that is more adaptable and efficient. The Three Climate Experiment provided a critical insight into the relationship between prior knowledge, data efficiency, and model performance. The results clearly

showed that the replay-enhanced approach allowed the model to retain valuable information from previous climates, enabling it to adapt more quickly and effectively to new environments. This adaptability is crucial for real-world applications where HVAC systems must operate under rapidly changing weather conditions.

Furthermore, the research highlighted the importance of explainability in DRL models. Engaging with building managers and industry professionals provided practical insights into how the model's decisions could be interpreted and applied in real-world settings. This interaction underscores the necessity of making advanced control algorithms not only powerful but also understandable and actionable for end users.

5.1.1 Contributions to knowledge

This thesis contributes new knowledge in several key areas:

1. **Three Climate Experiment Framework**: The novel framework we applied from RaE [4] for training and evaluating DDQNs across climates provides a new method for testing the generalisation capability of HVAC optimisation algorithms.

2. **Benefits of Replay Across Experiments (RaE) Methodology**: We empirically showed that incorporating the RaE methodology significantly enhances data efficiency and model robustness. The model could leverage knowledge from previous climate conditions to accelerate convergence and improve performance across both typical and extreme weather scenarios.

3. **Empirical Evidence of Adaptability to Climate Variability**: Through rigorous experimentation, we demonstrated that models trained with our framework can successfully adapt to modified and extreme climate conditions. This was validated using a novel dataset representing modern climate variability, proving the applicability of the approach in a changing global environment.

4. **Impact of Dataset Size on DRL Performance**: Our research demonstrated that larger datasets, especially with 100,000 transitions, lead to greater model stability, adaptability, and performance. This finding underscores the importance of prior data availability and size in improving DRL models' efficiency and performance for HVAC energy management.

5. **Explainability and Real-World Applicability**: We engaged with building managers to validate the explainability of our models in a real-world context. By identifying key areas of energy-saving potential, such as optimised control of central zones, we made the models' decision-making process interpretable and actionable for industry practitioners.

6. **Open-Sourced, Modular DRL Framework**: Finally, we refactored and open-sourced a modular and object-oriented DRL framework to facilitate further research and development. This modular approach enables other researchers to adapt and build upon the foundation laid in this thesis, contributing to faster iterations and innovations in energy management.

5.1.2 Recommendations for future work

While this research has made progress in optimising HVAC control, there are several areas where further work is needed:

- **Incorporation of Real-World Factors**: Future research should incorporate real building data as a prior. Also, it should work towards control of real buildings by encoding accessible state parameters and only using data as it is available in online deployment. This would make our solutions more applicable and valuable in practical settings.

- **Expansion to Other Building Models**: Testing the proposed framework across a broader range of building models and HVAC systems would enhance its gener-

alisability and applicability. This could include different building types, sizes, and HVAC configurations.

- **Advanced DRL Models**: Exploring more advanced DRL models, such as Rainbow DQN or Multi-Task Reinforcement Learning, could further improve performance, especially in complex and mixed-environment scenarios. These models could provide even greater adaptability and efficiency in HVAC control.
- **Explainability and User Interaction**: Continued focus on explainability is crucial for real-world deployment. Future work should explore ways to make the decision-making process of DRL models more transparent and user-friendly, allowing building managers to better understand and trust these systems.

5.1.3 Final remarks

In conclusion, this thesis has demonstrated that a replay-enhanced DRL approach can significantly improve HVAC control across diverse climates, achieving substantial energy savings while maintaining occupant comfort. The introduction of the RaE methodology and the Three Climate Experiment framework marks a significant new approach in the field, offering a robust, efficient, and generalisable solution to the complex challenge of HVAC optimisation in the built environment. Future work should build on these findings, exploring new models, incorporating real-world factors, and enhancing explainability to ensure that these advanced systems can be effectively deployed in practice.

References

[1] U. N. E. Programme, "Global status report for buildings and construction," tech. rep., March 2024.

[2] Z. Zhang and K. P. Lam, "Practical implementation and evaluation of deep reinforcement learning control for a radiant heating system," in *Proceedings of the 5th conference on systems for built environments*, pp. 148–157, 2018.

[3] Y. Du, H. Zandi, O. Kotevska, K. Kurte, J. Munk, K. Amasyali, E. Mckee, and F. Li, "Intelligent multi-zone residential hvac control strategy based on deep reinforcement learning," *Applied Energy*, vol. 281, p. 116117, 2021.

[4] D. Tirumala, T. Lampe, J. E. Chen, T. Haarnoja, S. Huang, G. Lever, B. Moran, T. Hertweck, L. Hasenclever, M. Riedmiller, N. Heess, and M. Wulfmeier, "Replay across experiments: A natural extension of off-policy RL," in *The Twelfth International Conference on Learning Representations*, 2024.

[5] H. Wang, X. Chen, N. Vital, E. Duffy, and A. Razi, "Energy optimization for hvac systems in multi-vav open offices: A deep reinforcement learning approach," *Applied Energy*, vol. 356, p. 122354, 2024.

[6] United Nations, "Transforming our world: the 2030 agenda for sustainable development," 2015.

[7] I. P. on Climate Change (IPCC), "Climate change 2023: Synthesis report," tech. rep., 2023.

[8] M. Kotz, A. Levermann, and L. Wenz, "The economic commitment of climate change," *Nature*, vol. 628, no. 8008, p. 551, 2024.

[9] W. M. Organization, "State of the global climate 2020 (wmo-no. 1264)," tech. rep., 2020.

[10] P. M. Forster *et al.*, "Indicators of global climate change 2022: annual update of large-scale indicators of the state of the climate system and human influence," *Earth System Science Data*, vol. 15, no. 6, pp. 2295–2327, 2023.

[11] A. Visioli, *Practical PID Control*. Springer-Verlag GmbH, 2006.

[12] K. S. Lee, I. S. Chin, H. J. Lee, and J. H. Lee, "Model predictive control technique combined with iterative learning for batch processes," *AIChE Journal*, vol. 45, no. 10, pp. 2175–2187, 1999.

[13] C. D. Corbin, G. P. Henze, and P. May-Ostendorp, "A model predictive control optimization environment for real-time commercial building application," *Journal of Building Performance Simulation*, vol. 6, no. 3, pp. 159–174, 2013.

[14] G. Mantovani and L. Ferrarini, "Temperature control of a commercial building with model predictive control techniques," *IEEE Transactions on Industrial Electronics*, vol. 62, no. 4, pp. 2651–2660, 2014.

[15] X. Hou, Y. Xiao, J. Cai, J. Hu, and J. E. Braun, "A distributed model predictive control approach for optimal coordination of multiple thermal zones in a large open space," 2016.

[16] J. Drgoňa, D. Picard, M. Kvasnica, and L. Helsen, "Approximate model predictive building control via machine learning," *Applied Energy*, vol. 218, pp. 199–216, 2018.

[17] Y. Yao and D. K. Shekhar, "State of the art review on model predictive control (mpc) in heating ventilation and air-conditioning (hvac) field," *Building and Environment*, vol. 200, p. 107952, 2021.

[18] T. Wei, Y. Wang, and Q. Zhu, "Deep reinforcement learning for building hvac control," in *Design Automation Conference (DAC)*, (Austin, TX, USA), pp. 1–6, 2017.

[19] T. Wei, S. Ren, and Q. Zhu, "Deep reinforcement learning for joint datacenter and hvac load control in distributed mixed-use buildings," *IEEE Transactions on Sustainable Computing*, vol. 6, no. 3, pp. 370–384, 2019.

[20] L. Yu, Y. Sun, Z. Xu, C. Shen, D. Yue, T. Jiang, and X. Guan, "Multi-agent deep reinforcement learning for hvac control in commercial buildings," *IEEE Transactions on Smart Grid*, vol. 12, no. 1, pp. 407–419, 2020.

[21] A. T. Nguyen, D. H. Pham, B. L. Oo, M. Santamouris, Y. Ahn, and B. T. Lim, "Modelling building hvac control strategies using a deep reinforcement learning approach," *Energy and Buildings*, vol. 310, p. 114065, 2024.

[22] Y. R. Yoon and H. J. Moon, "Performance based thermal comfort control (ptcc) using deep reinforcement learning for space cooling," *Energy and Buildings*, vol. 203, p. 109420, 2019.

[23] Z. Zou, X. Yu, and S. Ergan, "Towards optimal control of air handling units using deep reinforcement learning and recurrent neural network," *Building and Environment*, vol. 168, p. 106535, 2020.

[24] B. Chen, Z. Cai, and M. Berges, "Gnu-rl: A precocial reinforcement learning solution for building hvac control using a differentiable mpc policy," in *Proceedings of the 2019 ACM International Conference on Systems for Energy-Efficient Built Environments*, 2019.

[25] R. Džiugaitė-Tumėnienė, R. Mikučionienė, G. Streckienė, and J. Bielskus, "Development and analysis of a dynamic energy model of an office using a building management system (bms) and actual measurement data," *Energies*, vol. 14, no. 19, 2021.

[26] X. Liu and Z. Gou, "Occupant-centric hvac and window control: A reinforcement learning model for enhancing indoor thermal comfort and energy efficiency," *Building*

and Environment, vol. 250, p. 111197, 2024.

[27] B. Li and L. Xia, "A multi-grid reinforcement learning method for energy conservation and comfort of hvac in buildings," in *2015 IEEE International Conference on Automation Science and Engineering (CASE)*, (Gothenburg, Sweden), pp. 444–449, 2015.

[28] X. Fang, G. Gong, G. Li, L. Chun, P. Peng, W. Li, X. Shi, and X. Chen, "Deep reinforcement learning optimal control strategy for temperature setpoint real-time reset in multi-zone building hvac system," *Applied Thermal Engineering*, vol. 212, p. 118552, 2022.

[29] C. Lork, W.-T. Li, Y. Qin, Y. Zhou, C. Yuen, W. Tushar, and T. K. Saha, "An uncertainty-aware deep reinforcement learning framework for residential air conditioning energy management," *Applied Energy*, vol. 276, p. 115426, 2020.

[30] Z. Cheng, Q. Zhao, F. Wang, Y. Jiang, L. Xia, and J. Ding, "Satisfaction based q-learning for integrated lighting and blind control," *Energy and Buildings*, vol. 127, pp. 43–55, 2016.

[31] S. Qiu, Z. Li, Z. Li, J. Li, S. Long, and X. Li, "Model-free control method based on reinforcement learning for building cooling water systems: Validation by measured data-based simulation," *Energy and Buildings*, vol. 218, p. 110055, 2020.

[32] H. W. S. Yujiao Chen, Leslie K. Norford and A. Malkawi, "Optimal control of hvac and window systems for natural ventilation through reinforcement learning," *Energy and Buildings*, vol. 169, pp. 195–205, 2018.

[33] W. Valladares, M. Galindo, J. Gutiérrez, W.-C. Wu, K.-K. Liao, J.-C. Liao, K.-C. Lu, and C.-C. Wang, "Energy optimization associated with thermal comfort and indoor air control via a deep reinforcement learning algorithm," *Building and Environment*, vol. 155, pp. 105–117, 2019.

[34] S. Brandi, M. S. Piscitelli, M. Martellacci, and A. Capozzoli, "Deep reinforcement learning to optimise indoor temperature control and heating energy consumption in buildings," *Energy and Buildings*, vol. 224, p. 110225, 2020.

[35] X. Deng, Y. Zhang, Y. Zhang, and H. Qi, "Toward smart multizone hvac control by combining context-aware system and deep reinforcement learning," *IEEE Internet of Things Journal*, vol. 9, pp. 21010–21024, Nov. 2022.

[36] M. Vecerik, T. Hester, J. Scholz, F. Wang, O. Pietquin, B. Piot, N. Heess, T. Rothorl, T. Lampe, and M. Riedmiller, "Leveraging demonstrations for deep reinforcement learning on robotics problems with sparse rewards," 2017.

[37] A. Nair, B. McGrew, M. Andrychowicz, W. Zaremba, and P. Abbeel, "Overcoming exploration in reinforcement learning with demonstrations," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6292–6299, IEEE, 2018.

[38] A. Nair, M. Dalal, A. Gupta, and S. Levine, "Accelerating online reinforcement learning with offline datasets," 2020.

[39] A. Singh, A. Yu, J. Yang, J. Zhang, A. Kumar, and S. Levine, "Cog: Connecting new skills to past experience with offline reinforcement learning," *arXiv preprint arXiv:2010.14500*, 2020.

[40] L. Smith, J. C. Kew, T. Li, L. Luu, X. B. Peng, S. Ha, J. Tan, and S. Levine, "Learning and adapting agile locomotion skills by transferring experience," *ArXiv preprint arXiv:2304.09834*, vol. abs/2304.09834, 2023.

[41] H. Robbins and S. Monro, "A stochastic approximation method," *Annals of Mathematical Statistics*, vol. 22, no. 3, 1951.

[42] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: A Bradford Book, 2018.

[43] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[44] R. Bellman, *Dynamic Programming*. Princeton University Press, 1957. A Rand Corporation Research Study. Sixth Printing, 1972.

[45] R. S. Sutton, "Learning to predict by the methods of temporal differences," *Machine Learning*, vol. 3, no. 9, p. 44, 1988. Manufactured in The Netherlands.

[46] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Machine Learning*, vol. 8, no. 3-4, pp. 279–292, 1992.

[47] H. Hasselt, "Double q-learning," in *Advances in Neural Information Processing Systems 23 (NIPS 2010)*, 2010.

[48] R. American Society of Heating and A.-C. Engineers, "Ashrae handbook-fundamentals," 2021.

[49] E. Nikishin, M. Schwarzer, P. D'Oro, P.-L. Bacon, and A. Courville, "The primacy bias in deep reinforcement learning," in *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022.

[50] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1026–1034, 2015.

[51] M. Hessel, J. Modayil, H. Van Hasselt, T. Schaul, G. Ostrovski, W. Dabney, D. Horgan, B. Piot, M. G. Azar, and D. Silver, "Rainbow: Combining improvements in deep reinforcement learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[52] M. Fortunato, M. G. Azar, B. Piot, J. Menick, I. Osband, A. Graves, R. Munos, D. Hassabis, O. Pietquin, C. Blundell, *et al.*, "Noisy networks for exploration," in *International Conference on Learning Representations*, 2018.

[53] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized experience replay," in *International Conference on Learning Representations*, 2016.

[54] Z. Wang, T. Schaul, M. Hessel, H. v. Hasselt, M. Lanctot, and N. d. Freitas, "Dueling network architectures for deep reinforcement learning," in *International Conference on Machine Learning*, pp. 1995–2003, 2016.

[55] M. G. Bellemare, W. Dabney, and R. Munos, "A distributional perspective on reinforcement learning," *International Conference on Machine Learning*, pp. 449–458, 2017.

[56] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. P. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *International Conference on Machine Learning*, pp. 1928–1937, 2016.

A | Figures and Tables

Parameter	**Value**
timestep_per_hour	12
state_dim	15
action_dim	64
epochs	20
lr	0.001
gamma	0.9
epsilon	0.003
epsilon_min	0.001
epsilon_decay	0.95
target_update	200
buffer_size	10000
minimal_size	200
batch_size	128
T_factor_day	0.02
E_factor_day	1e-06
T_factor_night	0
E_factor_night	1e-06

Table A.1: DDQN Training Parameters

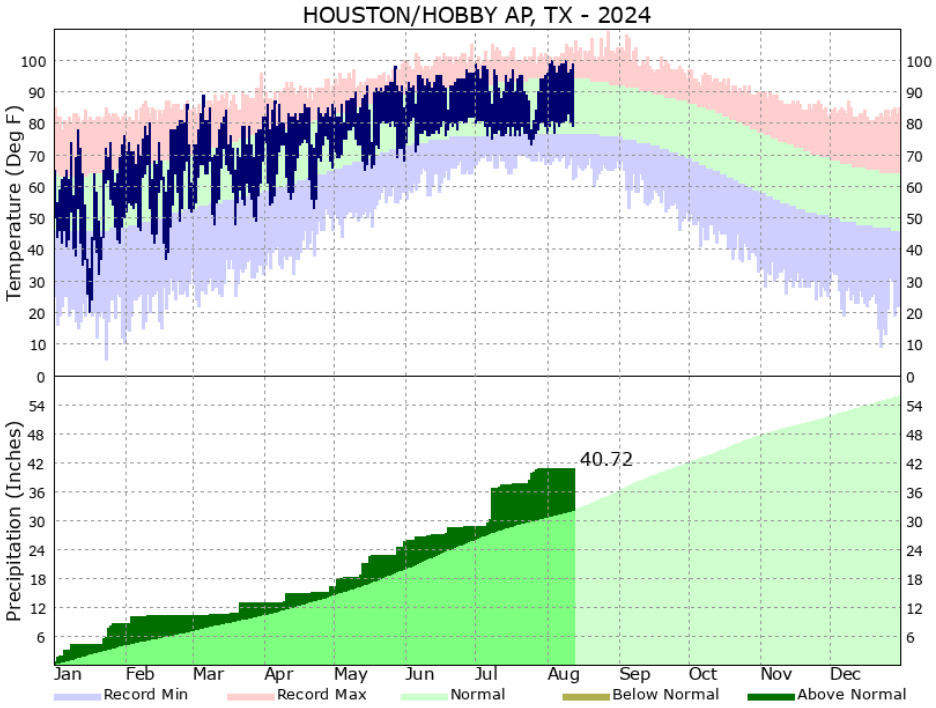


Figure A.1: Houston 2024 Weather

Country	**Location**	**Climate Zone**
Singapore	Changi (SIN)	Tropical Rainforest (Af)
UAE	Dubai Intl (DXB)	Hot Desert (BWh)
South Africa	Cape Town Intl (CPT)	Mediterranean (Csb)
Japan	Haneda (HND)	Humid Subtropical (Cfa)
Canada	Vancouver Intl (YVR)	Subarctic (Dfc)
USA	Santa Monica (SMO)	Mediterranean (Csa)
USA	Greenville-Spartanburg (GSP)	Humid Subtropical (Cfa)
USA	William P. Hobby (HOU)	Humid Subtropical (Cfa)
USA	Sky Harbor (PHX)	Hot Desert (BWh)
UK	London Heathrow (LHR)	Oceanic (Cfb)

Table A.2: Weather Files Description

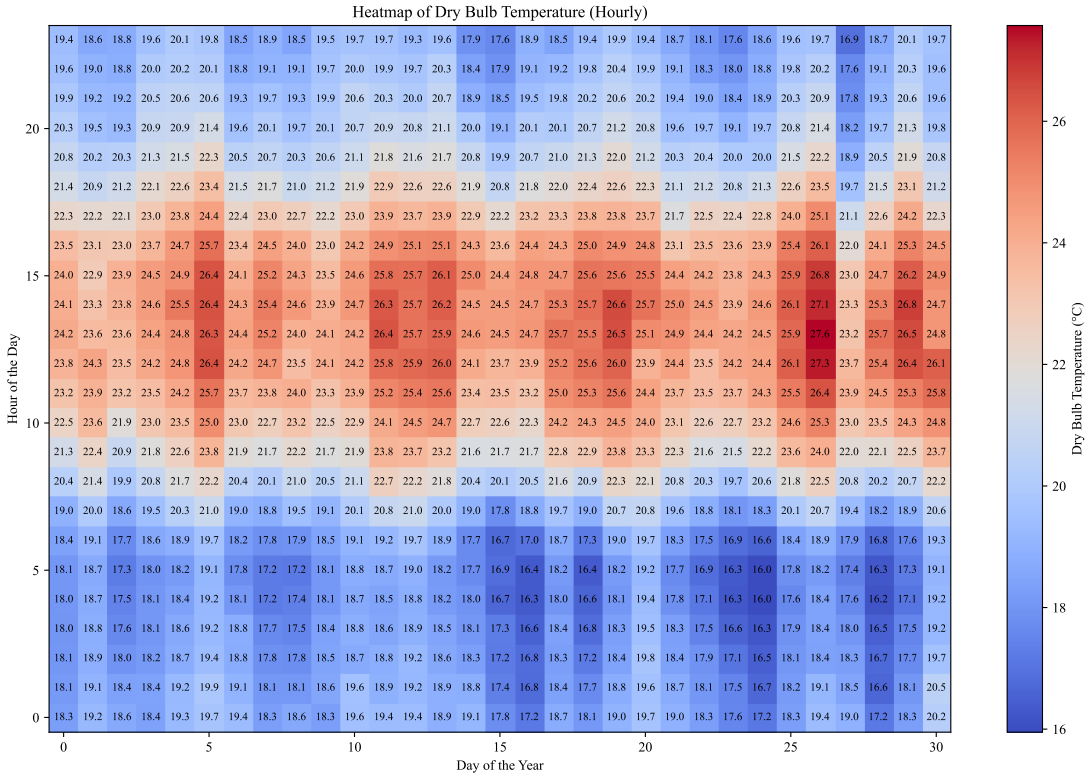


Figure A.2: Texas Hourly Temperatures (Default '.epw' file)

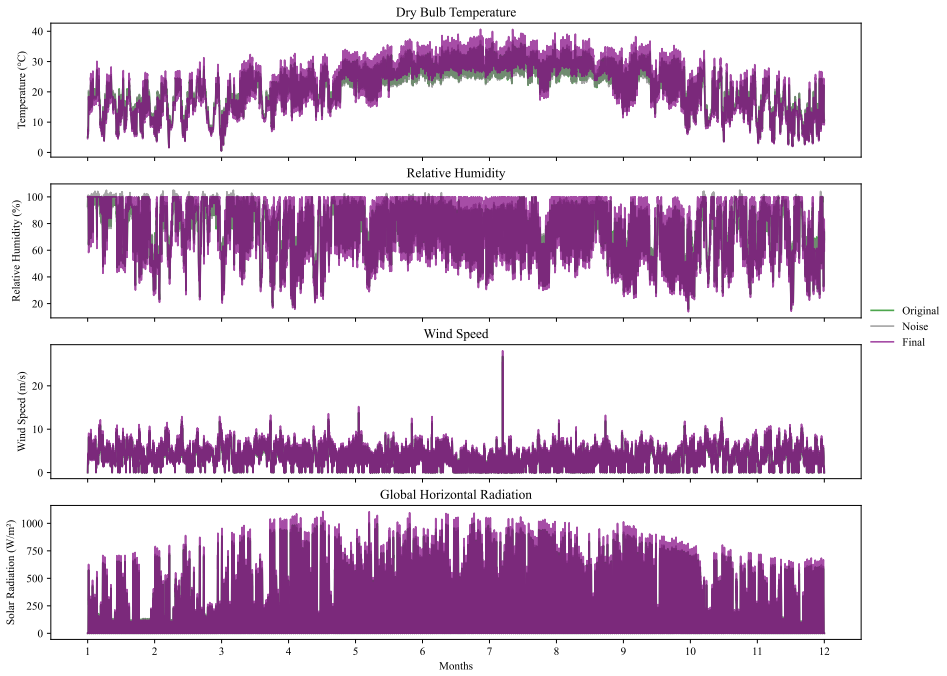


Figure A.3: Texas Weather File Modifications

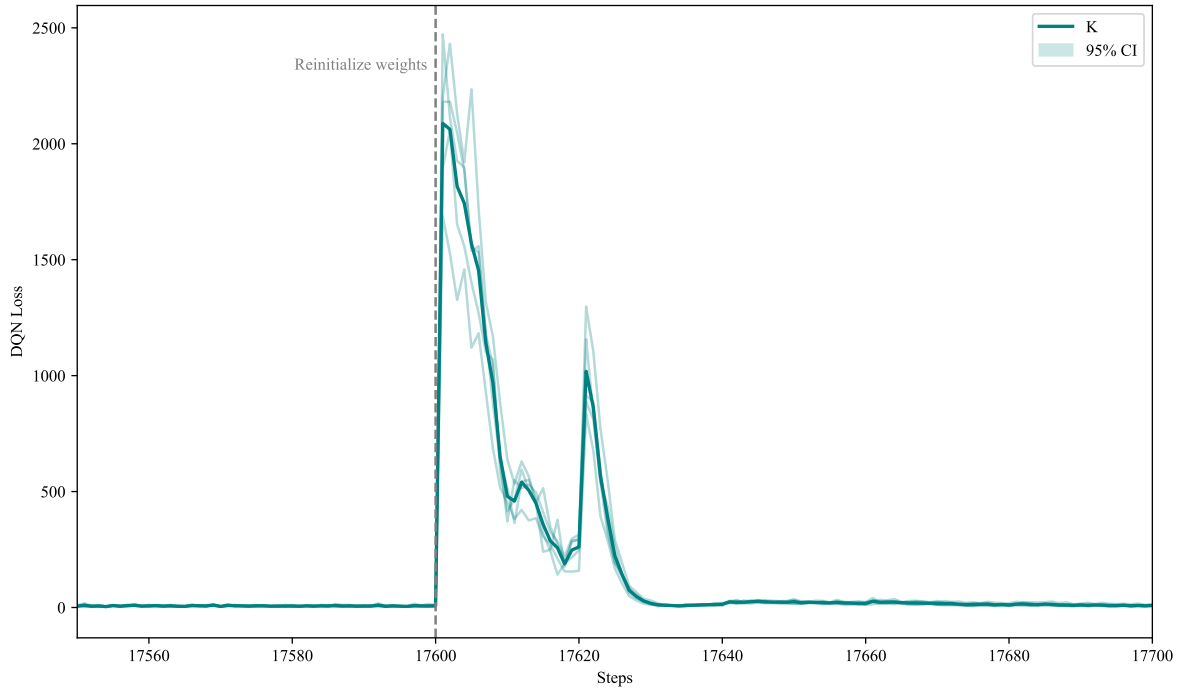


Figure A.4: Random Weight Resetting Directly After K (one) Re-initialisation

Climate	Reset Frequency	Average Reward	E_save_HVAC (%)	T_violation (%)	T_violation_offset	Work Time Length Ratio
ARIZONA	K	-1.7008	25.62	4.14	1.77	0.3577
	K/10	-1.7495	26.69	4.77	1.70	0.3577
	K/100	-1.8843	23.43	5.76	1.62	0.3577
CALIFORNIA	K	-0.9909	38.99	1.01	1.65	0.3576
	K/10	-1.0799	36.03	1.76	1.26	0.3576
	K/100	-1.1648	36.09	2.65	1.14	0.3576
DUBAI	K	-2.8820	11.11	8.74	2.53	0.3572
	K/10	-3.0247	9.71	9.93	2.40	0.3572
	K/100	-3.1429	8.69	10.83	2.35	0.3572
LONDON	K	-1.4193	36.28	1.35	1.57	0.3587
	K/10	-1.4291	37.61	1.90	1.35	0.3587
	K/100	-1.5973	37.68	4.09	1.27	0.3587
CAPE TOWN	K	-0.9863	41.51	0.94	1.83	0.3554
	K/10	-1.0028	42.77	1.36	1.42	0.3554
	K/100	-1.0512	43.09	2.08	1.25	0.3554
TEXAS	K	-1.7727	22.99	3.57	1.14	0.3575
	K/10	-1.8882	18.30	3.93	1.11	0.3575
	K/100	-1.9833	19.04	5.23	1.06	0.3575
SINGAPORE	K	-3.1148	7.52	11.50	0.89	0.3568
	K/10	-3.2121	6.97	12.18	0.97	0.3568
	K/100	-3.3615	3.56	12.99	1.02	0.3568
SOUTH CAROLINA	K	-1.6629	33.37	2.32	1.37	0.3577
	K/10	-1.7270	31.68	2.54	1.29	0.3577
	K/100	-1.8845	28.83	3.90	1.21	0.3577
TOKYO	K	-1.6727	28.20	2.64	1.17	0.3577
	K/10	-1.7785	25.58	3.62	1.09	0.3577
	K/100	-1.9119	23.68	4.81	1.21	0.3577
VANCOUVER	K	-1.4353	37.89	1.27	2.00	0.3585
	K/10	-1.4339	41.90	2.36	1.62	0.3585
	K/100	-1.6648	37.61	4.16	1.48	0.3585

Table A.3: Results from Random Weight Resetting Experiment (K, K/10, K/100)

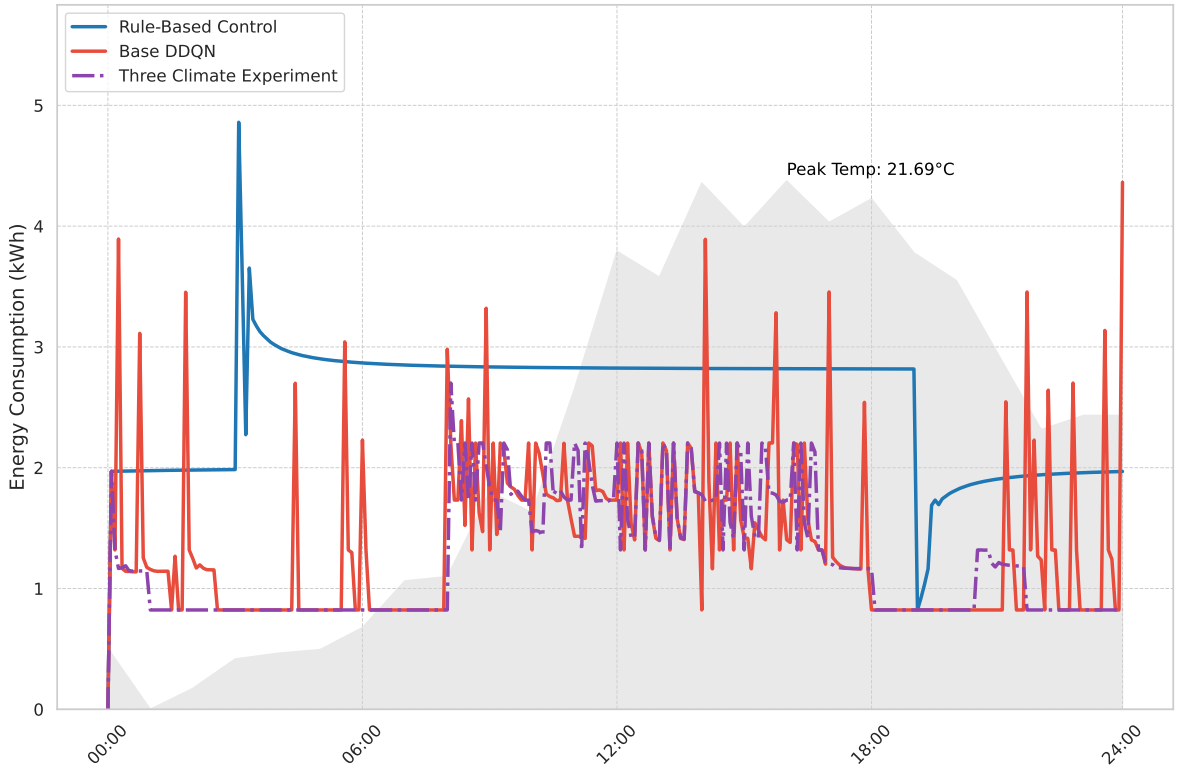


Figure A.5: Contrasting Base DDQN and Three Climate Experiment Control Behavior over a Sample Day

Location	**Climate**	**ES (%)**	**CV (%)**	**Δ ES Improvement**
Cape Town, SA	default	38.1917	0.5597	11.7115
Dubai, UAE	default	10.0881	8.1398	0.25863
London, UK	default	39.2745	2.04	-0.96098
Singapore, SG	default	5.6697	11.3084	-42.1198
Tokyo, JPN	default	25.6446	1.5729	7.52425
Vancouver, BC	default	39.8635	0.6508	-5.368824
California, USA	default	40.68492	1.49324	5.34208
Texas, USA	default	22.8663	2.9391	-1.479625
South Carolina, USA	default	29.6509	0.8472	7.986669
Arizona, USA	default	22.8821	2.6573	13.33548
Cape Town, SA	modified	38.6279	0.8614	6.05449
Dubai, UAE	modified	10.3175	12.3216	-3.2212
London, UK	modified	38.0588	2.3063	3.270715
Singapore, SG	modified	6.7453	15.9058	-30.1956
Tokyo, JPN	modified	22.6474	4.0528	3.252882
Vancouver, BC	modified	32.8970	0.8984	18.12672
California, USA	modified	36.2060	1.36612	3.2614
Texas, USA	modified	18.5431	6.5479	11.93328
South Carolina, USA	modified	26.6784	2.3499	6.9897
Arizona, USA	modified	22.1138	5.9350	-5.06207

Table A.4: Detailed Results of the Three Climate Experiment (**Standard** Dataset)

Location	Climate	ES (%)	CV (%)	Δ ES Improvement
Arizona, USA	default	62.57	2.94	53.02338
California, USA	default	69.67	1.04	34.32716
Dubai, UAE	default	55.41	8.20	45.58053
London, UK	default	69.52	1.31	29.28452
Cape Town, SA	default	70.15	0.99	43.6698
Texas, USA	default	61.32	3.26	36.974075
Singapore, SG	default	54.66	10.44	6.8705
South Carolina, USA	default	65.54	1.88	43.875769
Tokyo, JPN	default	62.95	2.14	44.82965
Vancouver, BC	default	70.17	1.83	24.937676
Arizona, USA	modified	60.27	7.18	33.09413
California, USA	modified	68.45	1.00	35.5054
Dubai, UAE	modified	55.05	12.88	41.5113
London, UK	modified	68.36	1.54	33.571915
Cape Town, SA	modified	69.49	1.08	36.91659
Texas, USA	modified	59.77	7.21	53.16018
Singapore, SG	modified	54.26	16.24	17.3191
South Carolina, USA	modified	63.63	2.59	43.9413
Tokyo, JPN	modified	60.84	4.90	41.445482
Vancouver, BC	modified	68.88	1.61	54.10972

Table A.5: Detailed Results of the Three Climate Experiment - (**Large** Dataset)

B | Interview Summary

Date: June 21

Time: 12:00 BST

Location: MS Teams

Participant: Philipp Mehrfeld, Recogizer

Summary: The participant discussed their experiences with managing HVAC systems in a campus setting, focusing on the challenges of data management, system optimisation, and the implementation of deep reinforcement learning (DRL) in smart buildings. The conversation also touched on the feasibility of optimising radiant cooling and the potential for applying DRL across different buildings.

- **Question 1: What are the critical data requirements for successful DRL implementation in the context of smart buildings?**

  The participant emphasised the importance of having comprehensive and well-organised data, particularly for Air Handling Units (AHU), static heating, and cooling circuits. However, they noted that the naming conventions and data points could be confusing, requiring a clear understanding of what each data point represents. They mentioned that while they have data for supply systems, indoor temperature sensors might not be fully covered.

- **Question 2: How do you ensure data quality and interpretability in your team?**

  The participant explained that data quality is a significant concern, particularly in understanding which data points correspond to specific aspects of the HVAC system. They have human-readable processed data available, but it's not consistently available for every room. The participant highlighted the challenge of processing data efficiently, which can take from a day to a couple of weeks, depending on the clarity of the naming conventions.

- **Key Insights:**
  - The participant's team has been working on optimising their campus's HVAC system using internally developed policies, including predictive algorithms based on weather forecasting.
  - They expressed a willingness to share a few days of data from specific buildings (BF5 and BF8) to assist with the research, although some data points may be missing or unclear.
  - The participant also discussed the challenges and potential of optimising radiant cooling systems, noting that while there is limited room for improvement in some cases, it depends on how the system behaves. They have not extensively explored transfer learning between different buildings, but they acknowledge that the buildings on their campus are quite similar, which could make this a viable approach.
  - The participant confirmed that their system is already controlled by algorithms and that they are open to exploring the integration of DRL for further optimisation. They also pointed out that the buildings do not track data on window usage, which could impact HVAC efficiency.