A Receding Horizon Reinforcement Learning Framework for Campus Chiller Energy Management - A case study from an Australian University

1st Laura Musgrave *
School of EECS
The University of Queensland
Brisbane, Australia
laura.musgrave@student.uq.edu.au

2nd Arnab Bhattacharjee *
School of EECS
The University of Queensland
Brisbane, Australia
a.bhattacharjee@uq.edu.au

3rd Tapan Kumar Saha
School of EECS
The University of Queensland
Brisbane, Australia
saha@eecs.uq.edu.au

***Abstract*—This work presents a case study of optimal energy management of a large Heating Ventilation and Cooling (HVAC) system within a university campus in Australia using Reinforcement Learning (RL). The HVAC system supplies to nine university buildings with an annual average electricity consumption of ~ 2 GWh. Updated chiller Coefficient of Performance (COP) curves are identified, and a predictive building cooling demand model is developed using historical data from the HVAC system. Based on these inputs, a Proximal Policy Optimization based RL model is trained to optimally schedule the chillers in a receding horizon control framework with a priority reward function for constraint satisfaction. Compared to the traditional way of controlling the HVAC system based on a reactive rule-based method, the proposed controller saves up to 28% of the electricity consumed by simply controlling the mass flow rates of the chiller banks and with minimal constraint violations.**

***Index Terms*—campus HVAC, chiller load management, campus energy management, deep reinforcement learning, model-free control**

I. INTRODUCTION

University campuses are large, interconnected energy systems where Heating, Ventilation, and Air Conditioning (HVAC) often accounts for more than 40% of total electricity consumption. Within HVAC systems, central chiller plants dominate energy use, typically contributing nearly 75% of HVAC electricity demand. Beyond improving energy efficiency, optimal chiller operation can support grid-level objectives. By modulating multi-megawatt cooling loads, campus plants can function as flexible resources capable of peak shaving, congestion relief, and ancillary service provision [1].

A typical campus chiller plant consists of multiple chillers with heterogeneous rated capacities and part-load characteristics, along with variable-frequency drive (VFD) pumps regulating chilled-water flow. These systems are intentionally over-provisioned to ensure redundancy across buildings. However, conventional rule-based or schedule-driven control strategies often result in part-load inefficiencies, with chillers operating away from their optimal efficiency regions. This can increase energy consumption by 10–20% compared to optimal coordination, motivating the development of intelligent multi-chiller scheduling strategies.

Chiller plant control approaches can be broadly categorized into rule-based, model-based, and data-driven methods. Rule-based control, commonly used in practice, relies on predefined load thresholds to sequentially switch chillers on or off [2]. While simple and robust, such approaches ignore heterogeneous part-load efficiencies and typically yield sub-optimal performance. Model-based methods such as Model Predictive Control (MPC) [3], [4] leverage predictive models of plant dynamics and cooling demand to compute optimal control actions. Although MPC remains the industry standard for constrained HVAC control, large multi-chiller plants require accurate modeling of nonlinear COP–PLR characteristics and explicit handling of binary on/off decisions. The resulting predictive problem is often formulated as a mixed-integer nonlinear program (MINLP), which can be computationally demanding and sensitive to model-plant mismatch in aging infrastructure. These challenges motivate exploration of model-light alternatives that retain predictive capability while reducing dependence on high-fidelity plant models.

Recent advances in data-driven control, particularly deep reinforcement learning (DRL), have demonstrated promise in learning optimal HVAC policies directly from operational data. Prior work includes LSTM-based load forecasting with Deep Q-Learning control [5], meta-learning-enhanced RL for improved transferability [6], and hybrid offline–online RL strategies for commercial cooling systems [7]. Despite these advances, most studies assume homogeneous chiller performance and rely primarily on historical load data, often neglecting exogenous weather variables. Furthermore, constraint handling remains a challenge, and reported energy savings typically plateau around 10–12%.

This work proposes a receding-horizon reinforcement learning framework for efficient operation of a four-chiller campus plant supplying nine buildings at a large Australian university. The proposed approach integrates predictive modeling and control to improve both energy efficiency and operational robustness. The main contributions are:

* Equal Contribution

979-8-3315-8138-1

1) **Receding-horizon RL control**: A Proximal Policy Optimization (PPO) agent [8] regulates chilled-water mass flow rates using 24-hour ahead load predictions in a receding-horizon structure.
2) **Heterogeneity-aware reward design**: A prioritized reward formulation promotes hard constraint satisfaction while incorporating distinct part-load efficiency characteristics of individual chillers.
3) **Weather-informed load forecasting**: A transformer-based TimeXer model generates multi-step cooling demand forecasts using historical loads and exogenous weather variables.
4) **Empirical benchmarking**: The proposed controller is compared against a practical rule-based baseline and a non-receding RL agent, achieving up to 28% energy savings relative to existing operation.

By combining domain-informed reward shaping, transformer-based forecasting, and receding-horizon RL, this study demonstrates a scalable and data-driven alternative for campus chiller optimization with potential relevance to grid-interactive energy management.

II. Problem Formulation

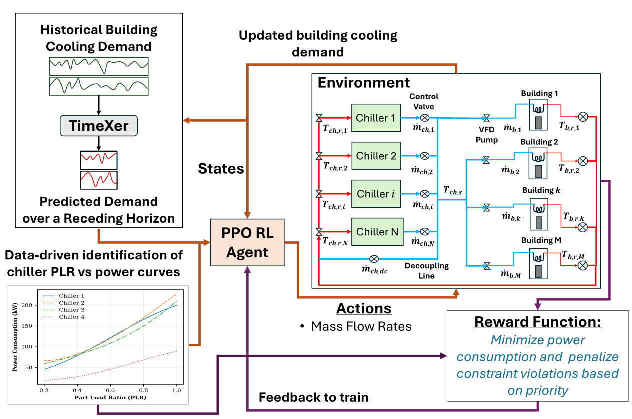

Fig. 1. A pictorial representation of a campus chiller network supplying to multiple buildings.

A receding horizon controller for predictive chiller load management is designed using a PPO-based RL agent. At every control instant, a building cooling load prediction algorithm generates possible building load trajectories over a 24 hour horizon which is used by the RL agent to schedule the chillers optimally by solving a constrained power minimization problem over a rolling horizon. To specify the differences in the part load efficiencies of the chillers, part-load ratio (PLR)-vs-Power consumption curves are pre-determined from historical data of the chillers and are explicitly incorporated in the reward function used to train the agent. An overview of the proposed RL framework is given in Fig.1 with the chiller network depicted as the environment. Specifically, the RL agent controls the mass flow rates of the chillers due to the following reasons:

1) Mass flow rates provide a wider range of control compared to supply temperatures and are controllable in the considered system through variable frequency motor systems.
2) The existing rule-based control strategy of the chillers fixes the supply temperature at $\sim 6^oC$ during working hours and $\sim 11^oC$ during non-working hours while varying the mass flow rates of the 'ON' chillers reactively based on load demand.

However, the authors acknowledge that this strategy may not be applicable in constant-flow chiller systems and will continue exploration of selective control of supply temperature and mass flow rates as part of future research.

A. Assumptions

In line with existing literature [3]–[6], the following assumptions were made when formulating the optimization problem - the reticulation pipework is lossless, the total mass flow rate from the chillers is equal to the total mass flow rate from the buildings, the supply water temperature for the chillers is fixed at 6^{o}C when the they are in operation, the return water temperature is equal for all chillers.

B. Objective Function

Consider a chiller bank with n_c chillers of varying cooling capacities $\dot{Q}_{r,i}$[1]. The intended goal is to set optimal mass flow rates for the chillers such that the power consumption of the chiller bank is minimized over a N step prediction horizon. This can be expressed mathematically by,

$$\min \sum_{t}^{t+N\cdot\Delta t} \left(\sum_{i=1}^{n_c} P_{c,i} \right) \tag{1}$$

A receding horizon approach is used, in which the optimization problem is solved after every Δt in the control horizon subject to the constraints outlined in the following sections.

C. Main Constraints

1) Power Consumption Equation: $P_{c,i}$ is the power consumed by chiller i over $[t, t+\Delta t]$, and can be represented by the equation,

$$P_{c,i} = \alpha_i + \beta_i \cdot PLR_i + \gamma_i \cdot PLR_i^2 + \psi_i \cdot PLR_i^3 \tag{2}$$

where $\alpha_i, \beta_i, \gamma_i, \psi_i$ are constant coefficients (see section IV-B for details), and PLR_i is the chiller Part Load Ratio.

2) PLR Equation: The PLR is the ratio of a chiller's cooling output $\dot{Q}_{c,i}$ to it's rated cooling capacity $\dot{Q}_{r,i}$, and is given by,

$$PLR_i = \frac{\dot{Q}_{c,i}}{\dot{Q}_{r,i}} \tag{3}$$

[1] $\dot{Q}$ denotes thermal power (kWr), where the dot represents the rate of heat exchange rather than a time derivative of energy.

979-8-3315-8138-1

3) Chiller Cooling Equation: The amount of cooling $\dot{Q}_{c,i}$ provided by chiller i is given by,

$$\dot{Q}_{c,i} = \delta_i \dot{m}_{c,i} c_w (T_{cr} - T_{cs}) \tag{4}$$

where $\dot{m}_{c,i}$ is the chiller mass flow rate, c_w is the specific heat capacity of water, T_{cr} and T_{cs} represent the return and supply water temperature respectively, and δ_i is a binary decision variable,

$$\delta_i = \begin{cases} 1, \text{ if chiller } i \text{ is } ON \\ 0, \text{ if chiller } i \text{ is } OFF \end{cases} \tag{5}$$

When a chiller is ON, it supplies chilled water at varying mass flow rates and a fixed supply temperature of approximately 6^{o}C. When OFF, it is considered shut down and supplies no chilled water. Thus, for a given chilled water return temperature T_{cr}, control of the cooling output (and by extension, the power consumption) of the chiller i can be achieved by varying it's mass flow rate $\dot{m}_{c,i}$.

4) Chilled Water Return Temperature: The temperature of the chilled water returning to the chiller bank T_{cr} varies according to the system dynamics (i.e. the heat retention characteristics of the building and the total thermal load). Given the assumption that the pipes are lossless, the return water temperature for the chillers can be expressed as,

$$T_{cr} = \frac{\dot{Q}_b}{\left(\sum_{i=1}^{n_c} \delta_i \dot{m}_{c,i}\right) c_w} + T_{cs} \tag{6}$$

5) Energy Balance Equation: The system energy balance is given by,

$$\sum_{i=1}^{n_c} \dot{Q}_{c,i} = \dot{Q}_b \tag{7}$$

6) PLR Operational Range: When chiller i is ON, it's PLR is subject to the constraint,

$$PLR_{min} \leq PLR_i \leq PLR_{max} \tag{8}$$

7) Return Temperature Operational Range: When any of the chillers are ON, the return temperature of the chiller bank is subject to the constraint,

$$T_{cr_{min}} \leq T_{cr} \leq T_{cr_{max}} \tag{9}$$

8) Minimum Operational Mass Flow Rate: When chiller i is ON, it's mass flow rate is subject to the constraint,

$$\dot{m}_{c,i_{min}} \leq \dot{m}_{c,i} \tag{10}$$

9) Maximum Coefficient of Performance: The Coefficient of Performance (COP) of chiller i is given by $COP_i = \frac{\dot{Q}_{c,i}}{P_{c,i}}$. The highest COP for chiller i is given by,

$$COP_i \leq COP_{i_{max}} \tag{11}$$

III. Optimal Control Strategy

A. Reinforcement Learning for Optimal Chiller Scheduling

A Proximal Policy Optimization based RL controller is proposed to learn a scheduling policy that selects the mass flow rates and, by extension, the ON/OFF status of the chillers.

1) State-space: The state vector s_t for the agent is given by,

$$s_t = \{\hat{\dot{Q}}_b, \dot{Q}_{c,1\ldots n_c}, PLR_{c,1\ldots n_c}, COP_{1\ldots n_c}, P_{c,1\ldots n_c}, T_{cr}\}_t \tag{12}$$

The building cooling load is given by predicted building cooling load, $\hat{\dot{Q}}_b$ in the state vector.

2) Action-space: The action space is given by,

$$a_t = \{m_{c,1\ldots n_c}\}_t \tag{13}$$

3) Reward: The reward function was developed with the aim of promoting constraint satisfaction while achieving the goal of minimizing power consumption. A sequential structure was used, implicitly assigning a priority to the constraints as the agent would receive a large negative reward until each constraint was met. Formally, suppose each constraint $C_i \forall i \in |\mathcal{C}|$ where $\mathcal{C}$ is the total constraint set, corresponds to a reward component r_i, then the reward function at step t is modelled as:

$$r_t = \sum_{c=1}^{|\mathcal{C}|} \lambda_c r_{c,t} \tag{14}$$

where $\sum_{c=1}^{|\mathcal{C}|} |\lambda_c| = max(\{|\lambda_c|\}_{c=1}^{|\mathcal{C}|})$ implies that any one of the reward components is active at time t. A pre-determined ordering strategy is used based on expert knowledge to select which constraints violations are more serious than others. Suppose, the priority is given by $f_P : \mathcal{C} \to \mathbb{N}$ such that $n = f_P(C_k)$ implies that the constraint C_k is the n^{th} constraint to be satisfied in order, then the following condition determines the value of the reward function at any time $t : |\lambda_k| > 0$ iff C_k is violated and $\forall i \in \mathcal{C}$ such that $f_P(C_i) < f_P(C_k)$, C_i is not violated. In this ordering, the reward component that corresponds to the power consumption is the least prioritized, ensuring that the agent focuses on meeting the hard constraints prior to minimizing power consumption. The hard constraints considered in this work are the energy balance equation (Eq. 7) and satisfying the return temperature bounds (Eq. 9) as they were found to be the hardest to satisfy during experiments. Only once the agent's action meets all these hard constraints can a positive reward be obtained. It should be noted that negative behaviours from the agent are penalized more heavily than positive behaviours are rewarded. The remaining constraints are considered soft constraints and included in the base reward with varying weights. Auxiliary soft constraints are included in the base reward to minimize frequent chiller switching and promoting sparsity in the number of operational chillers.

B. Building Cooling Load Forecasting

The receding horizon approach taken by the RL agent relies on the accurate forecasting of the building cooling load over the prediction horizon N. The TimeXer [9] module was selected for this task due to its ability to capture intra-endogenous temporal dependencies as well as exogenous-to-endogenous correlations in its forecasting approach. The exogenous variables used in the implementation of the forecaster include average wind speed and direction, average ambient

979-8-3315-8138-1

temperature, relative humidity, and average solar irradiance amount. The control time step is t = 30 minutes. The prediction horizon N = 48 therefore corresponds to a 24-hour ahead planning window (48 × 30 minutes). The forecasting model uses a look-back window of 336 time steps (7 days) to generate these 48-step predictions.

IV. Experiment Setup

A. Overall System Description

The chilled water refrigeration plant considered in this study is comprised of four water cooled chillers with three of them rated at 1700 kWr and one at 710 kWr, four cooling towers, chilled water and condenser water pumps, and associated pipework, valves and accessories. The upper limits of the mass flow rates for Chillers 1 to 3 and Chiller 4 are 50.8 L/s and 21.2 L/s respectively. This chilled water plant serves up to nine campus buildings at once via pipework which reticulates chilled water throughout the buildings to air handling units (AHUs), fan coils, and under-ceiling fan coils. The lower and upper limits for different parameters are as follows: $PLR_{min} = 0.2$, $PLR_{max} = 10$, $T_{cr,min} = 6.56$, $T_{cr,max} = 14$, $\dot{m}_{c,min} = 14.5$ and $COP_{max} = 10$ for Chillers 1 to 3 and $COP_{max} = 5.814$ for Chiller 4.

The RL agent utilizes the OpenAI SpinningUp PPO algorithm [10]. Training took approximately 3 hours using the Goolge Colab T4 GPU Runtime (maximum GPU RAM of 15.0 GB)

B. Chiller Power Consumption Model

Because the reward function incorporates chiller power consumption, a method of predicting this value for a given chiller cooling output was required. The degree three polynomial in Eq. 2 was used was fit to the historical data of the chillers to obtain the power curve coefficients $(\alpha_i, \beta_i, \gamma_i, \psi_i)$ and R^2 values for Chillers 1–4 as shown: C1 (33.3, -7.4, 384.9, -211.4; 0.93), C2 (78.6, -147.1, 454.0, -158.7; 0.87), C3 (39.5, 103.9, -53.5, 120.4; 0.91), and C4 (27.0, -115.1, 261.4, -84.1; N/A). Chiller 4 used the default mechanical specifications.

V. Results

A. Building Load Forecast

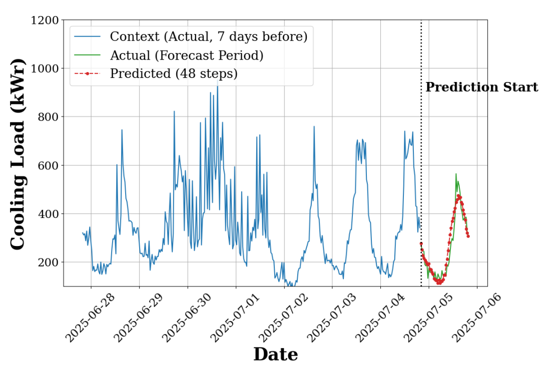

Fig. 2. Building load forecast with TimeXer

Fig. 2 shows the building forecast results using the TimeXer module for a specific testing instant. The normalized mean absolute error obtained for the testing period is 0.235, which results in quite accurate forecasts.

B. Benchmarking Results

TABLE I
Conceptual Comparison Between Classical MPC and Proposed RH-RL Framework

Feature	Classical MPC	Proposed RH-RL
Requires accurate plant model	Yes	No
Mixed-integer optimization	Often required	Implicit via policy
Explicit hard constraints	Yes	Reward-prioritized
Adaptation to model drift	Limited without re-identification	Possible via retraining
Computational burden	High for large-scale MINLP	Offline training; fast real-time inference

The training time return curve for the RL agent is shown in Fig. 3. The negative reward scale reflects the large penalty weights assigned to constraint violations; thus, convergence toward zero signifies increasing feasibility rather than direct physical energy interpretation. The power savings from using

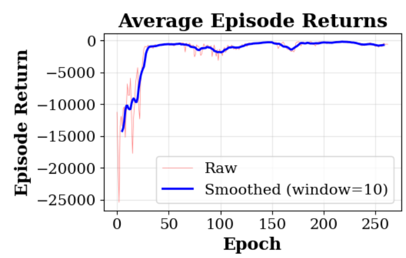

Fig. 3. Training convergence of the RL agent. The reward magnitude is dominated by penalties associated with high-priority hard constraint violations. Convergence toward a near-zero average episode return indicates that the agent has learned to satisfy hard constraints and is subsequently optimizing for energy efficiency.

the proposed RL is compared against two baselines[2] - the real-life rule-based controller and a pure RL controller that considers just the most recent power consumption in the state vector and does a one-step control instead of a receding horizon control. All three methods are evaluated over a two month evaluation period and the results are shown in Table II. The proposed RL-based receding horizon controller results in a total energy consumption of 66.49 MWh, compared to 92.35 MWh using the rule-based baseline and 74.35 MWh using the pure RL controller, resulting in a 28% savings in power consumption from the rule-based baseline applied in practice. These savings were achieved through two primary means: improving the accuracy of matching cooling demand and supply, and improving the COP of operating chillers.

C. Constraint Satisfaction

Fig. 4 compares the rule-based controller (blue) and the proposed RL agent (red). The RL controller achieves higher

[2]Although MPC is the industry standard, we don't include it as a baseline due to practical instabilities encountered during its implementation, largely originating from model-plant mismatches and the MINLP formulation. Instead we present a conceptual comparison between classical MPC and proposed RH-RL framework in Table I.

979-8-3315-8138-1

TABLE II
ENERGY SAVINGS COMPARISON ACROSS RULE BASED AND RL STRATEGIES. RH = RECEDING HORIZON

Control Method	Energy Consumed (MWh)	Energy Saved (MWh)	% Energy Saved
Rule-Based (in practice)	92.35	0	0
Pure RL (no RH)	74.15	18.2	19.7
Proposed RL	**66.49**	**25.86**	**28**

mean PLR (0.83 vs. 0.81) and COP (8.81 vs. 8.03), indicating improved operational efficiency. Temperature distributions remain within prescribed bounds for both approaches, while load-following accuracy is notably improved under RL control. The broader PLR dispersion reflects more effective utilization of heterogeneous chiller efficiency characteristics compared to sequential rule-based loading, contributing to the observed 28% energy savings.

Worst-case analysis over the two-month evaluation period shows that violations of the primary hard constraints—energy balance (Eq. 7) and return temperature limits (Eq. 9)—occurred in fewer than 5% of time steps. All violations were transient (2 consecutive intervals) and occurred only during very low cooling demand conditions ($<$ 300 kWr). The maximum deviation remained within 18% of baseline operational limits used in practice as safety bounds. No sustained infeasibility or cascading constraint breach was observed.

Although the RL controller does not provide formal feasibility guarantees as in classical MPC, the prioritized reward structure—with dominant penalties on hard constraint violations—empirically enforces bounded and infrequent deviations. Under the tested conditions, constraints therefore function as pseudo-hard constraints. Thermal comfort is preserved through strict energy balance enforcement, and no sustained unmet cooling demand was observed.

VI. CONCLUSION

This paper presented a receding-horizon reinforcement learning framework for optimal multi-chiller scheduling in a campus HVAC system. By integrating transformer-based load forecasting with a PPO agent and heterogeneity-aware reward design, the proposed controller achieves up to 28% energy savings while maintaining bounded and transient constraint violations. The current formulation assumes lossless pipework and relies on manually prioritized rewards. Future work will incorporate pipe losses, adaptive reward tuning, online learning mechanisms, and grid-interactive extensions to further enhance scalability and practical deployment. While the present study focuses on plant-level optimization, the receding-horizon RL framework readily extends to grid-interactive operation. Incorporating time-varying electricity prices, demand response signals, or feeder-level constraints into the reward function enables peak shaving, load shifting, and ancillary service participation. Since policy inference is computationally lightweight after offline training, the approach is compatible with day-ahead and intra-day energy management frameworks, directly linking campus HVAC flexibility with broader power system objectives.

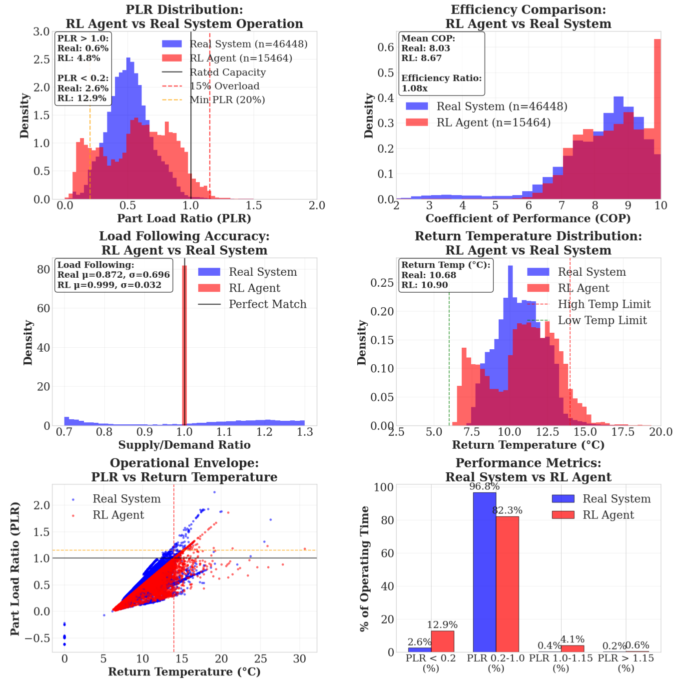

Fig. 4. Summary of results from the controller evaluation. Real System refers to the baseline rule-based controller currently in operation at the university

REFERENCES

[1] P. Yu, H. Zhang, Y. Song, H. Hui, and G. Chen, “District cooling system control for providing operating reserve based on safe deep reinforcement learning,” *IEEE Transactions on Power Systems*, vol. 39, no. 1, pp. 40–52, 2024.

[2] F. Jabari, M. Mohammadpourfard, and B. Mohammadi-ivatloo, “Energy efficient hourly scheduling of multi-chiller systems using imperialistic competitive algorithm,” *Computers Electrical Engineering*, vol. 82, p. 106550, 2020.

[3] Y. Ma, F. Borrelli, B. Hencey, B. Coffey, S. Bengea, and P. Haves, “Model predictive control for the operation of building cooling systems,” in *Proceedings of the 2010 American Control Conference*, pp. 5106–5111, 2010.

[4] B. Liu, Y. Gong, H. Luo, Y. Liu, B. Qin, and H. Li, “Model predictive control of hvac chilled water system based on load forecasting,” in *2024 6th International Conference on Intelligent Control, Measurement and Signal Processing (ICMSP)*, pp. 135–140, 2024.

[5] K. He, Q. Fu, Y. Lu, Y. Wang, J. Luo, H. Wu, and J. Chen, “Predictive control optimization of chiller plants based on deep reinforcement learning,” *Journal of Building Engineering*, vol. 76, p. 107158, 2023.

[6] P. Manoharan, M. P. Venkat, S. Nagarathinam, and A. Vasan, “Learn to chill: intelligent chiller scheduling using meta-learning and deep reinforcement learning,” in *Proceedings of the 8th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, BuildSys ’21, p. 21–30, ACM, Nov. 2021.

[7] J. Luo, C. Paduraru, O. Voicu, Y. Chervonyi, S. Munns, J. Li, C. Qian, P. Dutta, J. Q. Davis, N. Wu, *et al.*, “Controlling commercial cooling systems using reinforcement learning,” *arXiv preprint arXiv:2211.07357*, 2022.

[8] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal policy optimization algorithms,” *arXiv preprint arXiv:1707.06347*, 2017.

[9] Y. Wang, H. Wu, J. Dong, Y. Liu, Y. Qiu, H. Zhang, J. Wang, and M. Long, “Timexer: Empowering transformers for time series forecasting with exogenous variables,” *Advances in Neural Information Processing Systems*, 2024.

[10] J. Achiam *et al.*, “Spinning up in deep reinforcement learning.” https://spinningup.openai.com, 2018. Accessed: 2025-11-07.

979-8-3315-8138-1