Development of the Reward Function to support Model-Free Reinforcement Learning for a Heat Recovery Chiller System Optimization

Jean-Francois Landry[1], J. J. McArthur[1], Mikhail Genkin[1], Karim El Mokhtari[1,2]

[1] Toronto Metropolitan University (formerly Ryerson University), Toronto, Canada,
[2] FuseForward, Vancouver, Canada

jennifer.mcarthur@ryerson.ca

Abstract. Heat recovery chiller systems have significant strategic value to reduce building greenhouse gas emissions although this potential remains unrealized in practice. Real-time optimization using model-free reinforcement learning provides a potential solution to this challenge. A full-scale case study to implement reinforcement learning in a 6,000 m^2 academic laboratory is planned. This paper presents the methodology used to translate historical data correlations and expert input from operations personnel into the development of the reinforcement learning agent and associated reward function. This approach will permit a more stable and robust implementation of model-free reinforcement learning and the methodology presented will allow operator-identified constraints to be translated into reward functions more broadly, allowing for generalization to similar heat recovery chiller systems.

1. Introduction

Heat recovery chiller (HRC) systems permit simultaneous heating and cooling in buildings but their potential is unrealized in practice [1] as such systems often operate poorly or are not well understood, so they are often decommissioned by building operators [2]. Further, even when well-operated, optimization is extremely challenging due to the high degree of complexity required compared with standalone heating or cooling systems [3].

As we face the challenges of the climate crisis, buildings present a significant opportunity to decrease GHG emissions (IPCC, 2018) and HRC systems offer such savings, particularly where electricity generation relies on fossil fuels. This paper engages with the challenge of HRC optimization by exploring the supporting infrastructure and preliminary analysis necessary to develop and implement reinforcement learning (RL) in a full-scale facility.

Advances in reinforcement learning (RL) research to optimize HVAC applications demonstrate the high potential of the technology in building applications [4]. Compared with other ML techniques, RL is beneficial as it is an online learning method that can be implemented without a trained model [5]; this further permits its generalized application. To successfully implement RL, however, the appropriate reward function is critical. This paper presents insights on how to leverage operator expertise alongside historical data to develop this function, drawing from a full-scale HRC implementation at an academic facility.

2. Literature Review

To understand both the value of RL for HRC, we begin by providing a detailed overview of HRC operation drawing from both the literature and the system apparatus. Next, we present insights drawn from RL implementation in other heating and cooling optimization studies, discussing its merits, potential risks, and technical challenges to be overcome, which shaped our methodological approach

2.1. Heat Recovery Chiller Operation

HRC is a type of heat pump consisting of an evaporator section, which takes heat from the cooling system, and one or more condenser sections, which inject this heat into the heating system. When sized and operated correctly, these systems reduce the site energy use and improve load balancing [3]. A single condenser system [6] is the simplest form of HRC; in this system, the heat rejection from the cooling load is transferred to the heating loop via a heat exchanger providing pre-heating to a boiler loop. This allows heat rejection to lower temperatures to maximize this heat recovery. The HRC evaporator is installed in parallel with the chiller plant, which provides supplemental cooling as required. since HRC sizing is limited to the maximum cooling load where the heat rejection matches the base heat demand of the building [3].

The HRC operation is "confined" within an operating envelope [3] as follows. The HRC design capacity is the point at the intersection between the heating load and the heat rejection from the cooling load curves, as illustrated in Figure 1. At this point, a single control setpoint controls the HRC and both heating and cooling loads are equally prioritized. Below the associated outdoor air temperature (OAT), the heating system drops into a cooling-priority mode where the HRC load is driven by the evaporator leaving temperature (ELT) setpoint, dictated by the available cooling load for heat rejection. In such conditions, supplemental heating must be provided by the boiler system and the HRC condenser leaving temperature (CLT) setpoint has minimal effect. Similarly, above the design capacity OAT, the HRC is in heating-priority mode, driven by the CLT setpoint to meet the base heating load and supplemental cooling is required as the ELT setpoint has a minimum effect.

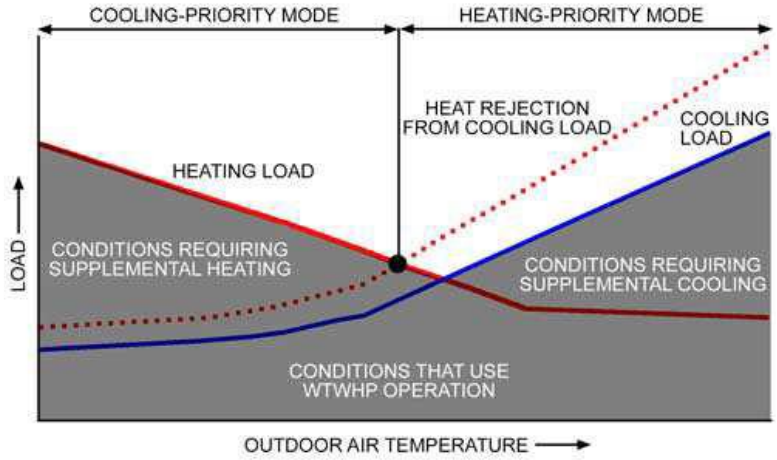


Figure 1. Selection and Operating Areas for Simultaneous Heating and Cooling HRC [3]

When this operating envelope is very large, HRC operation time is maximized but efficiency is suboptimal; when it is narrow, the converse is true. efficiency is maximized, but the operation time is sacrificed. HRC systems have many possible optimization parameters, such as the P_{hrc}, Qc, COP, the natural gas offset, the greenhouse gas emissions (GHG), and the cost savings (CS_{hrc}) [6]. Overall, the CS_{hrc} is likely the most appropriate aspect to minimize [1] [3] [7]. This is achievable by minimizing the compressor power (P_{hrc}) and maximizing the condenser side heat transfer (Q_c). Where natural gas savings are prioritized, HRC manufacturers indicate the importance of ELT and CLT setpoints; the

benefit of adjusting the CLT and increasing the F_{hrc} has been demonstrated in an academic study to result in savings of 43.9%, compared to the baseline [1].

2.2. Reinforcement Learning

RL is an area of Machine Learning (ML). ML is typically broadly sub-divided into three areas: 1 – supervised learning; 2 – unsupervised learning; 3 – RL. Supervised learning typically involves presenting an algorithm with labelled examples of data, that are used to train the underlying model. The model can then be used to make classification or regression predictions on un-labelled examples. Unsupervised learning involves discovering anomalies and patterns in the data rather than making predictions.

RL, on the other hand, involves training an intelligent agent to undertake a sequence of actions in a defined environment. The intelligent agent, once presented with a specific environment state, performs actions and collects either rewards or penalties. The intelligent agent seeks to maximize the total reward – defined by a mathematical function. RL offers significant generalization benefits compared with supervised ML techniques as it does not necessarily require a model for training, and therefore can optimize the system without prior knowledge [5].

2.3. Reinforcement Value of Reinforcement Learning for HVAC Optimization

Several studies have reviewed the application of RL within the building's context, summarized in [4]. Those specific to HVAC applications tended to use energy, flexibility, and comfort as the most commonly used control objectives while the most common RL algorithms were value iteration and tabular Q-learning, with a handful of studies instead using fitted Q-iteration or, wire fitted neural networks, or Fuzzy Q-learning. Two studies are of direct relevance to HRC optimization. On the chilled water side, [5] demonstrated the value of an RL model-free methodology to reduce energy use while maintaining the comfort of building occupants [5]. This study's results show that it was possible to converge an RL system over only in the summer season for a cooling system. On the heating side, [8] demonstrated the value of RL for predicting boiler combustion energy efficiency.

RL is valuable for HRC optimization as it mimics the behaviour of expert operators who could use trial and error to optimize the operating envelope to maximize cost savings; in practice, this is limited to a narrow range of operation whereas the RL could virtually all possible operating conditions. The cyclic nature of HRC operation provides repeated opportunities for this learning, benefiting rapid training of the RL agent. Third, the redundancy of the HRC with primary heating and cooling systems reduces the risk of RL experimentation. Finally, RL can be used to transfer of the knowledge learned from one system to another. This paper presents a method to supper this transferability to support HRC optimization at scale.

3. Research Methodology

The development of the RL implementation consisted of three steps: (1) analysis of the HRC system; (2) feature selection; and (3) RL agent development.

3.1. HRC Apparatus

A single condenser system was used as the test apparatus, shown schematically in Figure 2 and its design temperatures and rated performance data are summarized in Table 1. Real-time data is collected from this facility Building Automation System (BAS), including the condenser hydronic glycol flow (F_{hrcc}), and HRC entering (EET and CET for evaporator and condenser, respectively) and leaving (ELT and CLT) temperatures. To supplement BAS data, a dedicated power meter (Siemens MD-BM3; +/- 0.01kW accuracy) to measure the HRC's compressor power (P_{hrc}) and ultrasonic flow meter (Dynasonics Badger Meter TFX-5000; +/-0.001 l/s accuracy) were integrated with the BAS.

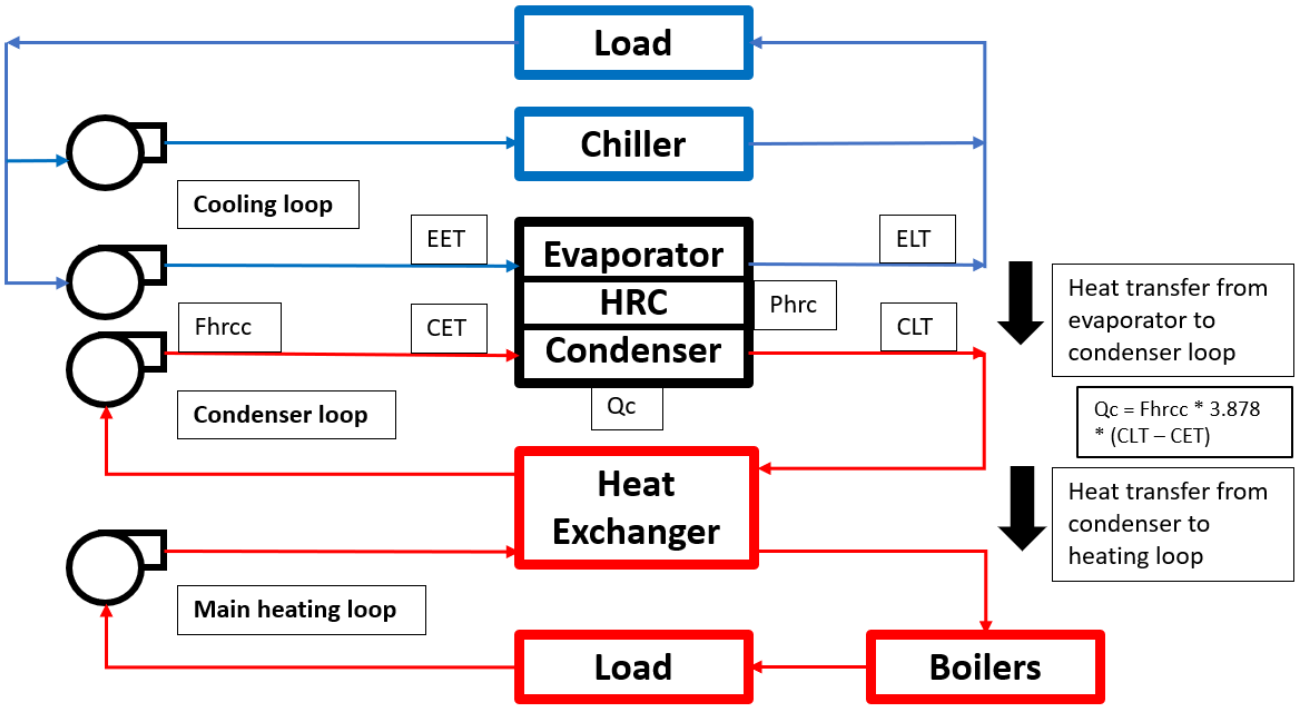


Figure 2. Single Condenser Heat Recovery Chiller System with variables indicated

Table 1. Heat Recovery AHRI and actual performance [9]

	Design Cooling Performance			*Evaporator*		*Condenser*		
Load	Capacity (kW)	Compressor (kW)	COP	Flow Rate l/s	ELT°C	Fhrc l/s	CET°C	CLT°C
100%	179.5	75.5	2.38	7.72	6.7	3.95	43.3	60.0
50%	89.7	38.0	2.36	7.72	6.7	3.95	51.7	59.4
	ACTUAL COOLING PERFORMANCE							
Per Comp	75.7	37.2	2.035	7.72	6.7	4.6	NA	60.0

3.2. Feature Selection

The desired features were identified based on the documentation of available data points, operator constraints, insights from the literature, and historical data analysis. Because the supplemental sensors were installed after the 2021 cooling season, only the heating season (OAT from -5°C to 17°C; cooling-priority) performance was evaluated. In addition to the measured variables indicated in Figure 2, two control variables – the ELT and CLT setpoints – were also considered as features. For simplicity, F_{hrcc} was maintained at 4.6 l/s by modulating the pump speed.

For the RL agent to take the best possible action, we need to determine the best candidate variables to represent the HRC's state. To achieve this goal, we proceeded by feature selection on 15 variables trended in the BAS. Highly correlated variables were removed using a correlation matrix. Then, we trained a linear regression model to predict the cost-saving (CS_{hrc}) from the remaining 13 variables after z-normalization. We removed the non-significant variables (p-value>0.05), then we eliminated sequentially less significant variables. The best linear regression model provided an R2 score of 0.994 with only 3 variables that are: P_{hrc}, CET and CLT. Therefore, the HRC's state will contain these variables in addition to the cost-saving (CS_{hrc}) and the number of operating compressors (Nb_{comp}).

3.3. Reward Function Development

The development of the agent was based on insights from the literature, experience of the facility engineers and operators gained during normal operations and commissioning, and data visualization. In typical building applications, operators would override setpoints based on their experience of the system and do adjustments depending on seasons or operational issues. The goal of the RL is to mimic this behavior but doing so perfectly. The RL agent interaction is shown schematically in Figure 3 and reads and acts on the HRC system state every minute.

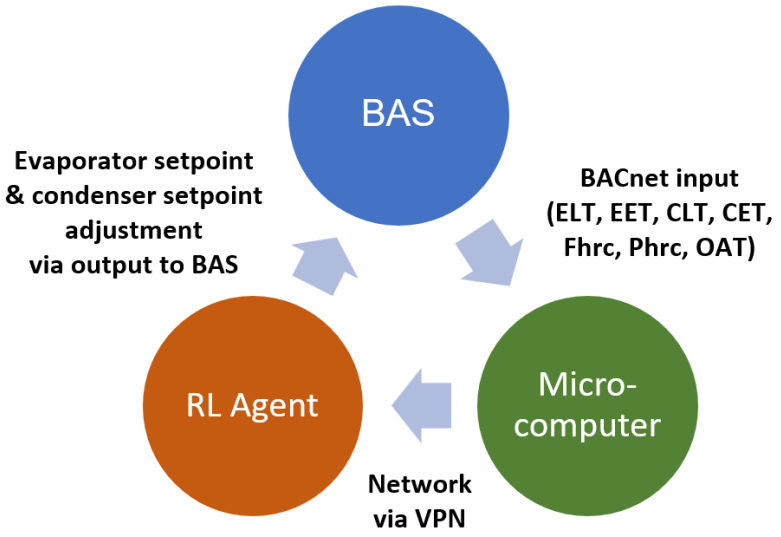


Figure 3. RL Agent interaction with the environment

Expert input facilitates efficient RL training while further minimizing risk. These were documented in the following operating constraints: (1) when the HRC is operating, at $Nb_{comp} \geq 1$, the savings are maximized; (2) when the HRC is operating, at $Nb_{comp} = 0$, no savings are obtained and heat can migrate from the main heating loop to the condenser loop, creating energy losses. (3) when the HRC is operating, cycling from $Nb_{comp} = 2$ to $Nb_{comp} = 0$ compressors repeatedly indicates unstable operations and creates an unnecessary strain on the HRC and an unsteady operation.

4. Reward Function Development

4.1. Objective

The goal of the RL is to find the best operating envelope at all times, that will maintain the HRC in operation in the most efficient way, therefore maximizing the cost savings. Equation 1 presents the hourly cost savings calculation for HRC operation compared to baseline operation (subscript b). To consider both GHG and energy impacts, carbon taxes have been included in energy costs.

$$CShrc = ((\text{Nbcomp} * \text{Phrc}_b) - \text{Phrc}) * \left(cost_{\text{electricity}}\right) + (\text{Qc} - (\text{Nbcomp} * \text{Qc}_b)) * \left(cost_{naturalgas}\right) \quad (1)$$

This cost savings is the target value for optimization and was selected to consider the balance between energy (P_{hrc} and Q_c) and the GHG reduction; to achieve this, the system lift (defined as CLT-ELT) needs to be minimized [3] [7].

4.2. Insights from Data Analysis

Data analysis highlighted four concerns. First, when the CLT setpoint was too low (operating envelope too narrow), the chiller would cycle directly from 0 to 2 compressors as illustrated in Figure 4. This unstable operation could result in premature failure and must be avoided by using a harsh RL penalty for 0 chiller operation. Expert knowledge learned from the data. The data acquired gives an insight into the desired operation of the RL agent below the baseline.

Second, the HRC pumps continue even when the HRC is not operating, resulting in a migration of heat from the boiler loop to the HRC loop and negative system capacity. Finally, the transient performance resulted in noise and improbable values in the measured data, aggravated by BAS-induced data lags.

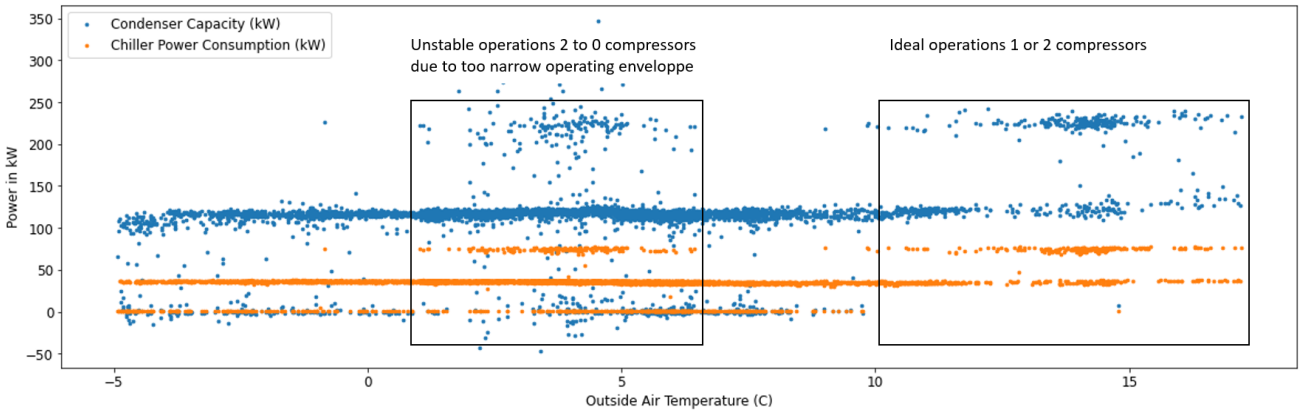
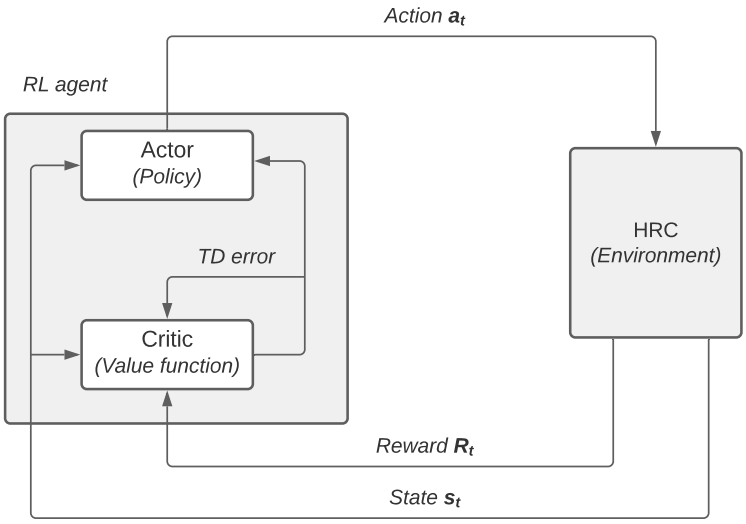


Figure 4. Distributions and co-distributions of the pre-processed training data

4.2.1. Architecture & pseudo code.

In the RL architecture depicted in Figure 5, the HRC is the environment that operates under an unknown model. It is observable through a state st reported every 30 seconds and controlled by the action at provided by the RL agent. A reward Rt is calculated from the environment and used by the RL agent to gradually improve its policy. The selection of the RL agent architecture is dictated by various considerations such as how much the agent knows about the environment's dynamics, whether the interaction is episodic or not, and the nature of the action (discrete or continuous). In the HRC case, the model is unknown, and the actions are continuous (evaporator and condenser setpoints). Many stable architectures can be applied in this case such as the Advantage actor-critic A2C on Figure 5 [10] [11], DDPG [12] or SAC [13]. While the role of the actor's network is to find the optimal policy by acting on the HRC (action at), the critic's network evaluates the policy produced by the actor and provides feedback (TD error) to guide both networks to maximize the reward R_t in subsequent interactions (see Table 2).



$$a_t = \begin{pmatrix} Evaporator\_STPT \\ Condenser\_STPT \end{pmatrix}$$

$$s_t = \begin{pmatrix} Nb_{comp}^t \\ CS_{hrc}^t \\ P_{hrc}^t \\ CET^t \\ CLT^t \end{pmatrix}$$

Figure 5. RL agent architecture, adapted from [10, 11]

RL agent nomenclature

ϕ	Critic's network parameter	s_t	HRC state at instant t
θ	Actor's network parameter	s_{t+1}	HRC state at instant t+1
π_θ	Actor's policy	a_t	Agent's action at instant t
J(.)	Neural network loss function	a_{t+1}	Agent's action at instant t +1
∇	Gradient operator	R_t	Reward obtained by the agent at instant t

The reward function is given by the following equation:

$$R_t = \epsilon\, \rho_t CS_{hrc}^t + (CS_{hrc}^t - CS_{hrc}^{t-1}) \tag{2}$$

$$\text{with } \rho_t = \begin{cases} 0 & \text{if } Nb_{comp}^t = 0 \\ 1 & \text{otherwise} \end{cases}$$

The total reward R_t is a linear combination of two terms. The first term is a product of the learning rate (ϵ), the limited number of operating compressors at time t (ρ_t), and the total cost savings calculated at time t (CS_{hrc}^t) to ensure that any actions resulting in cost savings receive positive reinforcement.

The term ρ_t which can be described as the limited number of operating compressors, can have two possible values. It is equal to 0 if the actual total number of operating compressors (Nb_{comp}^t) is also equal to zero. It is limited to the maximum value of 1 in those cases when the actual number of operating compressors (Nb_{comp}^t) is greater than or equal to 1. This is done to reflect the fact that operating more than 1 compressor does not typically result in additional cost savings. Furthermore, allowing ρ_t to take on values greater than 1 would end up giving the first term of the reward function equation too much weight, making it less sensitive to changes in cost savings imparted by the most recent action.

It was observed that when 0 compressors are operating, no cost savings are possible due to lack of heat exchange. The product of the limited number of compressors operating (ρ_t) and the total instantaneous cost savings (CS_{hrc}^t) describes this effect. If the total number of operating compressors drops to 0 ($Nb_{comp}^t = 0$), the first term of the reward function equation will be eliminated. If one or more compressors are operating, then this term will significantly influence the total value of the reward.

The learning rate hyperparameter (ϵ) will be typically set to a value between 0 and 1. It is used to control the extent to which the first term of the reward function equation influences the overall reward and prevents it from becoming too dominant in those situations when more than one compressor is operating. Small values of ϵ will make the reward function more sensitive to the effects of the latest action. Larger values of ϵ will make the reward function less sensitive to the effects of the latest action. Reducing sensitivity can help mitigate algorithm oscillation.

The second term in the reward function equation is the difference between the current instantaneous cost savings calculated at time t (CS_{hrc}^{t}), and the previous instantaneous cost savings calculated at time t-1 (CS_{hrc}^{t-1}). In those situations when the current cost savings are lower than previous cost savings the agent will receive a lower reward. In those situations when the current cost savings are close to zero, or much lower than the previous cost savings, the second term of the equation will dominate the sum and will result in a negative reward being returned to the agent.

Table 2. Actor-Critic's algorithm pseudo-code overview

```
Initialize actor and critic network parameters θ and ϕ respectively
for each step t:
  Select an action in the state s_t using the actor's policy: a_t~π_θ(s_t)
  Take action a_t and receive reward R_t, move to next state s_{t+1}
  Compute policy gradient for the actor: ∇_θ J(θ)
  Update the actor's network parameter θ using gradient descent
  Compute the loss of the critic's network J(ϕ)
  Compute the gradient for the critic ∇_ϕ J(ϕ)
  Update the critic's network parameter ϕ using gradient ascent
end for
```

The resultant action and reward table is summarized in Table 3.

Table 3. RL actions and rewards

Condition	Potential Actions	Reward	Constraints
CShrc $_{s_{t+1}}$ < CShrc $_{s_t}$	Do nothing ELT Setpoint +0.56°C CLT Setpoint −0.56°C	Negative reward	4.4°C ≤ ELT ≤ 8.9°C 48.9°C ≤ CLT ≤ 60°C CLT ≥ HWT-SP+5°C; ELT and CLT setpoints may only be adjusted in increments of 1°F (0.56°C)
CShrc $_{s_{t+1}}$ = CShrc $_{s_t}$	Do nothing ELT Setpoint +0.56°C CLT Setpoint −0.56°C	No reward	
CShrc $_{s_{t+1}}$ > CShrc $_{s_t}$	Do nothing	Positive reward	
Nbcomp $_{s_{t+1}}$ = 0 and Nbcomp $_{s_t}$ = 1	ELT Setpoint -0.56°C CLT Setpoint +0.56°C	Large negative reward	
Nbcomp $_{s_{t+1}}$ = 0 and Nbcomp $_{s_t}$ = 0	ELT Setpoint -0.56°C CLT Setpoint +0.56°C	No reward	
Nbcomp $_{s_{t+1}}$ = 0 and Nbcomp $_{s_t}$ = 2	ELT Setpoint +/- 0.56°C CLT Setpoint +/− 0.56°C	Very large negative reward	

5. Discussion & Conclusion

Using a model-free RL method for the HRC system is promising, especially for the HRC. Research on HVAC optimization is growing and this article helps to grow the knowledge-based around this subject. This paper presents the development of a reward function as informed by operator expert input and presents a generalized method to replicated this for other HRC systems, updated with system-specific constraints and operational parameters.

5.1. Limitations

The context of data collection is the most significant limitation of this study. The case study data are acquired after the Covid-19 Pandemic, but yet, the occupancy is not typical to pre-pandemic, since building offices remain mostly vacant, with many non-lab employees working remotely. This has an impact on both the heating and cooling building load in office spaces. The HRC operations were disturbed from time to time due to maintenance as well, or planned shutdown on the cooling and heating system. Further, the additional sensor data only captured the cooling-priority mode and partial data at design condition (two compressors); having the HPM condition is necessary to establish the complete baseline for savings calculation baseline. Second, the transient operation of the HRC resulted in noisy and unexpected data and signal processing is required to ensure the RL agent is trained on appropriate data. In addition, two simplifications were used in the preliminary model. First, we did not evaluate the potential benefits of a variable F_{hrcc} to further optimize the chiller operation because the manufacturer recommends a constant F_{hrcc} for this machine. Second, only the average energy rate is considered in this study, to simplify the calculation. This means the fees and demand rate are blended and their segregation could further increase cost savings.

5.2. Future Research

Implementation of the RL agent is tentatively scheduled for early spring of 2022. During this implementation, a parametric analysis will be completed to understand the impact of the cost of electricity, natural gas, and the carbon tax on the optimization of the operating setpoint. RL learning will also be evaluated to adjust or refine the reward function. The installation of a glycol reserve tank on the condenser side could also be used to further improve performance, serving as a damper to reduce the cycling process, further increasing the energy reduction from the system and controllability.

6. Acknowledgements

This research has been funded by the Natural Science and Engineering Research Council (NSERC) Alliance Grant (ALLRP- 544569-19) and FuseForward.

References

[1] Wang L, Sakurai Y, Bowman SJ and Claridge DE 2018 Commissioning an existing heat recovery chiller system. *ASHRAE Journal*, pp. 44-52

[2] Durkin T and Rishel JB 2003 Dedicated Heat Recovery. *ASHRAE Journal,* **45(10)**, pp. 18-22

[3] Campbell CR, Catrambone JA and Paraskevakos CP 2012 Large-capacity, water-to-water heat pumps for centralized plants. *ASHRAE Journa,l* **54(5)**, p.26

[4] Wang Z and Hong T 2020 Reinforcement Learning for Building Controls: The opportunities and challenges. *Applied Energy,* **269**, p.115036

[5] Qiu S, Li Z, Fan D, He R, Dai X and Li Z 2022 Chilled water temperature resetting using model-free reinforcement learning: Engineering application. *Energy and Buildings,* **255**, p. 111694

[6] Dorgan CB, Dorgan CE and Linder RJ 1999 *Chiller heat recovery application guide*. American Society of Heating Refrigerating and Air-Conditioning Engineers

[7] Heemer J, Mitrovic A and Scheer M 2011 Increasing central plant efficiency via a water to water heat pump. *Pharmaceutical Engineering,* **31(3)**, pp.1-8

[8] Jiang H, Cai Z, Zhang T and Peng C 2021 Prediction of boiler combustion energy efficiency via deep reinforcement learning. *40th Chinese Control Conference (CCC)*:IEEE, pp. 2658-2662

[9] Multistack, 2017 *Multistack Dedicated Heat Recovery Chiller Submittal for Approval – Revision 1.* Multistack

[10] Mnih V, Badia AP, Mirza M, Graves A, Lillicrap T, Harley T, Silver D and Kavukcuoglu K 2016 Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*: PMLR, pp. 1928-1937

[11] Grondman I, Busoniu L, Lopes GA and Babuska R 2012 A survey of actor-critic reinforcement learning: standard and natural policy gradients. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews),* **42(16)**, pp. 1291-1307

[12] Lillicrap TP, Hunt JJ, Pritzel A, Heess N, Erez T, Tassa Y, Silver D and Wierstra D 2015 Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*

[13] Haarnoja T, Zhou A, Abbeel P and Levine S 2018 Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*: PMLR, pp.1861-1870