# Chapter 150
# Model-Free Reinforcement Learning-Based Control for Radiant Floor Heating Systems


**Xu Han and Ali Malkawi**

**Abstract** This paper explores the feasibility and strategies of using model-free reinforcement learning-based control (RLC) for the slow response radiant floor heating (RFH) systems with a setback setting. First, a detailed physics-based virtual testbed is developed and validated. Then based on the virtual testbed, four different strategies of RLC to handle the slow response are studied, along with a conventional rule-based control (RBC) without setback as a baseline and an MPC with a setback for the upper bound on the performance. The results show that the DQN_TD(λ) with forecasted weather data as states provides the best performance, showing potential for applications. Compared to the baseline, the heating demand is reduced by 19.1% with RLC and 18.5% with MPC. The unmet hours of RLC with our settings are higher than that of MPC, which suggests that more research is needed for RLC to better meet the constraints.

**Keywords** Reinforcement learning based control · Model predictive control · Radiant floor heating

## 150.1 Introduction

Radiant floor heating (RFH) systems have been demonstrated to provide better thermal comfort while reducing energy consumption. However, the control of such systems is challenging because of the high thermal inertia and corresponding slow response time (Arteconi et al. 2014). Rule-based control (RBC) has been demonstrated to be effective in continuous operation by controlling the supply water temperature or concrete core temperature as a function of outdoor weather conditions, which can be the mean ambient temperature of the past 24 h (Kalz 2010) or weather forecast of the ambient temperature and solar radiation (Hoogmartens and Maarten 2011).

X. Han (✉) · A. Malkawi
Graduate School of Design, Harvard University, Cambridge, MA 02138, USA
e-mail: xu.han@gsd.harvard.edu

Center for Green Buildings and Cities, Harvard University, Cambridge, MA 02138, USA



L. L. Wang et al. (eds.), *Proceedings of the 5th International Conference on Building Energy and Environment*, Environmental Science and Engineering,
https://doi.org/10.1007/978-981-19-9822-5_150

However, the RBC may become problematic when a temperature setback during an occupied period at night is applied, which may cause delayed heating during the following morning.

Model predictive control (MPC) has been proposed to address this issue. Two studies demonstrated that 17–24% (Privara et al. 2011) and 19–25% (Gayeski et al. 2012) of energy were saved by using MPC. However, the necessity of an accurate dynamic model limits its wide applications in real practice. For example, a grey-box model for the thermal dynamics was developed in Hoogmartens and Maarten (2011), which requires expert knowledge. In contrast, as a model-free algorithm, reinforcement learning (RL) is promising to provide a practical solution based on a few attempts that have been made for HVAC control. However, reinforcement learning-based control (RLC) for the RFH systems has not been thoroughly studied.

Zhang and Lam (2018) applied a deep RL control method to a radiant heating system, which was found to be able to save 16.6–18.2% of energy compared to an RBC. However, they pointed out that future work should focus on the delayed reward problem caused by a slow thermal response. And, other optimal control methods (e.g., MPC) need to be included to further evaluate and benchmark the effectiveness of RLC. Blad et al. (2019) utilized two different RL algorithms with Q-networks to control an underfloor heating system and found that adding eligibility trace showed a better performance than standalone Q-networks for cases with slow dynamics but performed slightly worse for cases without slow dynamics, which revealed the potential of eligibility trace to deal with slow dynamics. However, the environment is simplified and hypothetical without validation using measured data, which did not include solar radiation, windows, real weather data and setback. Also, as mentioned in the paper, they did not take forecasted weather data into account in the RL model, which may have improved the performance. Arroyo et al. (2022) proposed a reinforced model-predictive control algorithm and tested it with a floor heating system. They found that the proposed algorithm can meet constraints and achieve similar performance as MPC. However, this algorithm requires system identification and is consequently not a pure model-free method.

This paper explores the feasibility and strategies of using model-free RLC for the slow response RFH systems with a setback setting. To benchmark the effectiveness of the proposed RLC strategies, a conventional rule-based control (RBC) without setback is chosen as a baseline and an optimal MPC with a setback is included to provide the upper bound on the performance. A detailed physics-based virtual testbed for a single zone is developed in Modelica and validated based on operating data of a real building.

## 150.2 Methods

The section introduces the proposed four RLC strategies along with RBC and MPC. First, the RBC, serving as a baseline, is designed based on the control logic running in the real system. Then, in the MPC, a state-space model is developed with the

simulation data from the RBC case, and an MPC optimization problem is formulated. In the RLC, four different strategies are proposed. Finally, the development and validation of the physics-based virtual testbed are described.

### *150.2.1 Rule-Based Control (RBC)*

A commonly used RBC for a radiant floor heating system is to control the slab temperature instead of the zone air temperature. To deal with the high thermal inertia, the setpoint of the slab temperature is modulated as a function of the weather forecasts. The zone air temperature is therefore maintained relatively stable within an acceptable range taking advantage of the self-regulating effect. Refer to Yan et al. (2022) for more detailed descriptions of the RBC. The weighted forecasted outdoor air dry-bulb temperature $\overline{T}_E$ is defined as average of hourly outdoor air temperatures for the next 24 h. The setpoint of the slab temperature ($T_m$) is calculated with:

$$T_m = 27.8 - 0.18 \times \overline{T}_E, \tag{150.1}$$

The slab temperature is controlled by modulating the flowrate and temperature of the water supplied to the radiant systems of the room. The flowrate is controlled with an on/off heat valve. The supply water temperature ($T_{sup}$) is determined by:

$$T_{sup} = 43.6 - 0.76 \times \overline{T}_E, \tag{150.2}$$

### *150.2.2 Model Predictive Control (MPC)*

A linear state-space model is identified with a discrete formulation:

$$x(t+1) = Ax(t) + Bu(t) + Fd(t) + K, \tag{150.3}$$

where, $x(t+1)$ and $x(t)$ are the states and response variables at the time step of $t+1$ and $t$, respectively, which denote the room air temperature $T_r$ in this study. $u(t)$ denotes the vector of control inputs, on/off of heat valve $u_{val}$, and supply water temperature $u_{Tsup}$. $d(t)$ denotes the vector of disturbances, which in this study is presence of occupancy $D_{occ}$, outdoor dry bulb temperature $D_{Tdb}$, and solar radiation $D_{sol}$. $A$, $B$, $F$ and $K$ are system matrices that need to be parameterized based on the input and output data. After identifying the state-space model, an MPC optimization problem is formulated:

$$\min \sum_{k=0}^{N-1} \delta^{T_r}(k) + w_1 \cdot \sum_{k=0}^{N-1} u_{val}(k) + w_2 \cdot \sum_{k=1}^{N-1} (u_{val}(k) - (u_{val}(k-1))^2 \quad (150.4)$$

$$s.t. \quad T_r(k+1) = a_1 T_r(k) + \begin{bmatrix} b_1 & b_2 \end{bmatrix} \begin{bmatrix} u_{val}(k) \\ u_{Tsup}(k) \end{bmatrix} + \begin{bmatrix} f_1 & f_2 & f_3 \end{bmatrix} \begin{bmatrix} \hat{D}_{occ}(k) \\ \hat{D}_{Tdb}(k) \\ \hat{D}_{sol}(k) \end{bmatrix} + K, \quad k = 0, \ldots N-1, \quad (150.5)$$

$$\underline{T}_r(k) - \delta^{T_r}(k) \le T_r(k) \le \overline{T}_r(k) + \delta^{T_r}(k), k = 1, \ldots N, \quad (150.6)$$

$$\underline{T}_r(k) = \begin{cases} 15 \; if \; \hat{D}_{occ}(k) = 0 \\ 20 \; if \; \hat{D}_{occ}(k) = 1 \end{cases}, k = 1, \ldots N, \quad (150.7)$$

$$\overline{T}_r(k) = \begin{cases} 32 \; if \; \hat{D}_{occ}(k) = 0 \\ 26 \; if \; \hat{D}_{occ}(k) = 1 \end{cases}, k = 1, \ldots N, \quad (150.8)$$

$$\delta^{T_r}(k) \ge 0, k = 1, \ldots N, \quad (150.9)$$

$$35 \le u_{Tsup}(k) \le 55, k = 0, \ldots N-1, \quad (150.10)$$

$$u_{val}(k) \in 0, 1\}, k = 0, \ldots N-1, \quad (150.11)$$

$$\hat{D}_{occ}(k) = D_{occ}(k), k = 0, \ldots N-1, \quad (150.12)$$

$$\hat{D}_{Tdb}(k) = D_{Tdb}(k) + \omega_{Tdb}, \omega_{Tdb} \in N\left(0, 0.3^2\right), k = 0, \ldots N-1, \quad (150.13)$$

$$\hat{D}_{sol}(k) = D_{sol}(k) + \omega_{sol}, \omega_{sol} \in N\left(0, 5\%^2\right), k = 0, \ldots N-1, \quad (150.14)$$

In Eq. 150.4, the three terms to minimize are the discomfort ($\delta^{T_r}$ denoting the deviation of $T_r$ out of the comfort range), total opening period of heat valve (energy), and frequent cycling of heat valve, where $w_1$ and $w_2$ are weighting factors (0.02 and 0.05, respectively, in this study). Equation 150.5 is the identified dynamic system. Equations 150.7and150.8 are the bounding constraints for room temperature for thermal comfort considering a setback on occupied hours (lower setpoint at night to save energy). Equations 150.9–150.11 are bounding constraints of three optimization variables. In Eq. 150.12, a perfect prediction of the occupancy status is assumed. In Eqs. 150.13and150.14, the predictions of outdoor dry bulb temperature and solar radiation are assumed to be adding a Gaussian noise to the ground truth.

### *150.2.3 Reinforcement Learning-Based Control (RLC)*

The RL algorithm learns an optimal policy $\pi: S_t \rightarrow A_t$ that maximizes the accumulated future rewards $\sum_t^{\infty} R_t$ through interactions between the agent and the environment ($S_t$, $A_t$, $S_{t+1}$, $R_{t+1}$). Temporal difference methods including TD(0) and TD(λ) are used to update the model in this study. To train the RL model, TD(0) uses information from only one step ahead to perform an update, which is biased. In contrast to TD(0), the Monte-Carlo method is not biased but needs to wait until the end of a complete episode to perform an update. The FHS is continuously running without a clear concept of "episode" and therefore may not be suitable for this method. TD(λ), also known as Eligibility Traces method, extends TD(0) to perform an update from n-steps look-ahead, which shows potential to deal with slow and delayed responses in the floor heating systems (Blad et al. 2019). Therefore, TD(λ) is adopted in this study with TD(0) as a benchmark. The Deep Q-Network (DQN) is used in this study with memory reply for improving training efficiency. The loss function of the DQN with TD(0) can be expressed as:

$$L(\theta) = \mathbb{E}_{(\hat{s},a,r,\hat{s}')\sim U(D)}\left[\left(r + \gamma \max_{a'\in\mathcal{A}} Q\left(\hat{s}', a'; \theta^{-}\right) - Q\left(\hat{s}, a'; \theta\right)\right)^2\right]. \quad (150.15)$$

With the n-step returns involved in the TD(λ), a new loss function becomes:

$$L(\theta) = \mathbb{E}_{(\hat{s},a,r,\hat{s}')\sim U(D)}\left[R^{\lambda} - \left(Q\left(\hat{s}, a'; \theta\right)\right)^2\right], \quad (150.16)$$

The detailed calculations of $R^{\lambda}$ can be referred to Daley and Amato (2019). Four cases are proposed to investigate the performance of different learning strategies, i.e., TD(0) and TD(λ), and state designs, as shown in Table 150.1. The third and fourth cases take the weather forecast information as states to potentially improve the ability of the RLC to deal with the slow response FHS.

**Table 150.1** State design for different cases

| Cases | States (all standardized) |
|---|---|
| DQN_TD(0) | Room temperatures at current and last time steps, orientation (room temperature compared to setpoint), outdoor dry-bulb temperature, solar radiation, occupancy, hour of the day, heat valve status |
| DQN_TD(λ) | |
| DQN_TD(0) + WF[a] | In addition to the states in DQN_TD(0), more states are added, including predicted averaged outdoor dry-bulb temperature and solar radiation in the next several hours |
| DQN_TD(λ) + WF | |

[a] *WF* weather forecast

**Table 150.2** Results of calibration and validation of the virtual testbed

| Outputs | Calibration | | Validation | |
|---|---|---|---|---|
| | MAE | RMSE | MAE | RMSE |
| $T_{slab}$ | 0.71 | 0.91 | 0.57 | 0.83 |
| $T_{returnWater}$ | 0.71 | 0.93 | 0.69 | 0.85 |
| $T_{room}$ | 0.54 | 0.73 | 0.65 | 0.76 |

### *150.2.4 Virtual Testbed*

This study assesses the performance of different control algorithms based on a virtual testbed developed with Modelica and Python (Wetter et al. 2014). The modeled room is a customizable laboratory with floor heating systems on the third floor of the small office building and living laboratory, located in Cambridge, Massachusetts, named HouseZero (Yan et al. 2022). The room model is calibrated and validated with measured data from the third-floor lab of HouseZero. The measured data in December 2020 is divided into two parts, in which the data from December 1st to 15th is used for calibration and the data from December 16th to 31st is used for validation. The results are shown in Table 150.2. The errors are less than 1 K for all the parameters for both calibration and validation.

## 150.3 Results

The training of the RLC algorithms is performed with the weather data from 2021 November to 2022 January and repeated as five episodes (three months per episode), leading to a total training period of 920 days and total iterations of 88,320 with a time step of 15 min. The RLC algorithms are tested with the weather data of February after the training. The training curves of the reward score for the four RLC strategies over episodes are shown in Fig. 150.1. The two algorithms with TD(0) show relatively lower reward scores while the other with TD(λ) and the weather forecast shows the best overall performance, in which the step in TD(λ) is set to be 48 (corresponding to 12 h) and 1-12-h' weather forecasts are adopted as states. A further tunning of these parameters is performed, and it is found that the step of 60 and 1-24-h' weather forecasts are the optimal settings to achieve the best performance, which is taken to compare with the baseline and MPC.

Table 150.3 shows the comparisons of the heating demands (energy perspective) and unmet hours (comfort perspective) with RBC, MPC and RLC. Compared to baseline (RBC), RLC achieves a similar level of energy savings (19.1%) as MPC (18.5%). This is achieved by avoiding overheating of the zone and adopting a setback at night through control improvements without any retrofits of the systems. However, the unmet hours have not been improved in the RLC case against the baseline, while

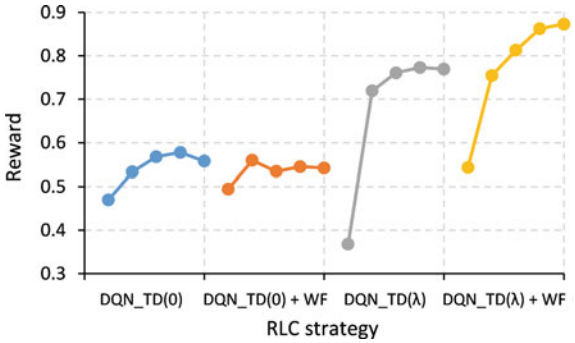


**Fig. 150.1** Training curve of the reward score over episodes with different RLC strategies

**Table 150.3** Total heating demands and unmet hours in February with different controllers

| Cases | Heating demand (kWh) | | | Savings (%) | Unmet hours (h) | | |
|---|---|---|---|---|---|---|---|
| | Occupied period | Unoccupied period | Total | | Occupied period | Unoccupied period | Total |
| RBC | 102.2 | 160.9 | 263.1 | – | 19 | 0 | 19 |
| MPC | 26.7 | 187.6 | 214.3 | 18.5 | 7 | 1 | 8 |
| RLC | 52.5 | 160.3 | 212.8 | 19.1 | 19 | 0 | 19 |

it is reduced from 19 h in baseline to 8 h in the MPC case. Most of the unmet hours occur during an occupied period.

A more detailed analysis regarding the room temperature control performance with different controllers is shown in Fig. 150.2. For the occupied period, the temperatures range from 18.8 to 25.4 °C with RBC, from 19.7 to 23.8 °C with MPC, and from 19.4 to 24.3 °C with a few outliers below 19.4 °C with RLC, while the comfort range is defined from 20 to 26 °C in this study. For the unoccupied period, the temperatures range from 18.2 to 23.7 °C with RBC, from 15.0 to 23.2 °C with MPC, and from 15.9 to 21.3 °C with a few outliers over 21.3 °C with RLC, while the acceptable range is defined from 15 to 32 °C. It is found that the MPC and RLC maintain the room temperatures more precisely within the comfort range during the occupied period while the room temperatures are controlled to be at a much lower level. This is still in the acceptable range during the unoccupied period at night, which may explain the energy savings through the two control algorithms. In most of the unmet hours in both the MPC and RLC cases, the room temperatures are only slightly lower than the lower bound.

The detailed control dynamics with different controllers in a selected period from February 10th to 15th is shown in Fig. 150.3. The first plot shows the room temperature control trajectories. Both MPC and RLC achieve good temperature control during the day and a setback at night to save energy, in which the trajectory in RLC mostly aligns with that in MPC but slightly less optimal. For example, in RLC, the temperature is a little lower than 20 °C at 9 am on February 10th, and the temperature at night does not drop to a minimum level to maximize the energy savings as

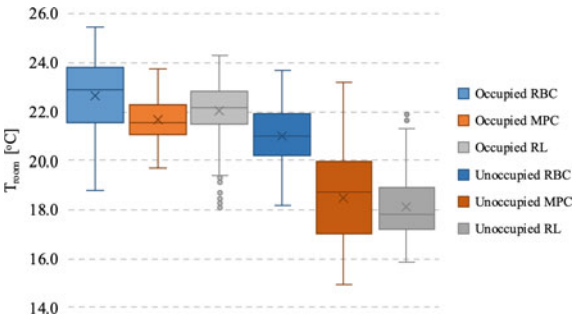


**Fig. 150.2** Statistical distribution of room temperatures with different controllers in February

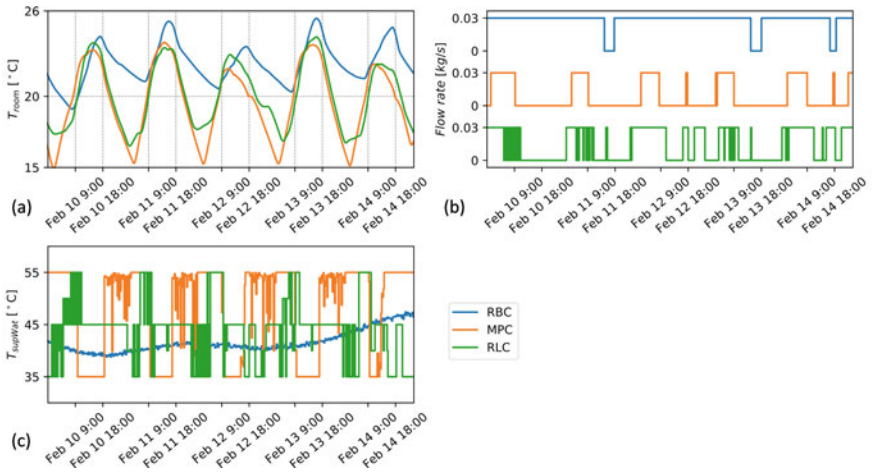


**Fig. 150.3** **a** Room air temperature, **b** supply water flow rate and **c** supply water temperature with different controllers from February 10th to 15th

MPC. The second and third plots show the control sequences determined by different controllers. The model-free RLC generally learns a control policy that is similar as MPC to operate the FHS in an intermittent manner though still not so optimal as MPC. The fourth plot is the weather data in the studied period.

## 150.4 Discussion and Conclusions

This paper proposed a new strategy, i.e., DQN_TD(λ) with weather forecasts, to apply model-free RLC to control FHS with a slow response. The effectiveness of the proposed strategy was demonstrated and benchmarked by comparing with RBC as a baseline, MPC as an upper bound of the performance, and other RLC strategies

without TD(λ) or weather forecasts. The assessment was conducted in a physics-based virtual testbed that was validated with measured data from a real building. The results showed that the proposed strategy achieved a similar level of energy savings (19.1%) against the baseline as MPC (18.5%). Meanwhile, the results of more unmet hours occurring with RLC than that in MPC reveal that the constraints were not strictly met in RLC though the room temperatures are only slightly lower than the lower bound in most of the unmet hours. This suggests more research is needed to optimize the RLC to address the constraint violation issue. In summary, the proposed strategy of model-free RLC shows comparable performance against MPC to solve the challenging control problem for slow response RFH systems with a setback setting. Though all the findings are subject to studied cases in this paper with our implemented models, the proposed strategy demonstrated its potential to be applied in other buildings by adopting a similar model and training the model in a new environment. Future research will include improving the performance and testing the generalization performance of the RLC algorithm with other cases as well as in real building systems.

**Acknowledgements** The authors are grateful for the help from Runyu Zhang and Na Li from SEAS Harvard.

## References

Arroyo J, Manna C, Spiessens F, Helsen L (2022) Reinforced model predictive control (RL-MPC) for building energy management. Appl Energy 309:118346

Arteconi A, Costola D, Hoes P, Hensen J (2014) Analysis of control strategies for thermally activated building systems under demand side management mechanisms. Energy Build 80:384–393

Blad C, Koch S, Ganeswarathas S, Kallesøe C, Bøgh S (2019) Control of hvac-systems with slow thermodynamic using reinforcement learning. Procedia Manuf 38:1308–1315

Daley B, Amato C (2019) Reconciling λ-returns with experience replay. Adv Neural Inf Process Syst:32

Gayeski N, Armstrong P, Norford L (2012) Predictive pre-cooling of thermo-active building systems with low-lift chillers. HVAC&R Res 18(5):858–873

Hoogmartens J, Sourbron M (2011) Review report of existing control strategies for GEO-HP-TABS

Kalz DE (2010) Heating and cooling concepts employing environmental energy and thermo-active building systems for low-energy buildings: system analysis and optimization, Ph.D. Fakultät für Architektur der Universität Karlsruhe

Privara S, Široký J, Ferkl L, Cigler J (2011) Model predictive control of a building heating system: the first experience. Energy Build 43(2–3):564–572

Wetter M, Zuo W, Nouidui TS, Pang X (2014) Modelica buildings library. J Build Perform Simul 7(4):253–270

Yan B et al (2022) Comprehensive assessment of operational performance of coupled natural ventilation and thermally active building system via an extensive sensor network. Energy Build:111921

Zhang Z, Lam KP (2018) 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