Towards Optimal District Heating Temperature Control in China with Deep Reinforcement Learning

Adrien Le Coz
EDF R&D China Center
Beijing, China
adrien.le-coz@edf.fr

Tahar Nabil
EDF R&D China Center
Beijing, China
tahar-t.nabil@edf.fr

François Courtot
EDF R&D China Center
Beijing, China
francois.courtot@edf.fr

Abstract

Achieving efficiency gains in Chinese district heating networks, thereby reducing their carbon footprint, requires new optimal control methods going beyond current industry tools. Focusing on the secondary network, we propose a data-driven deep reinforcement learning (DRL) approach to address this task. We build a recurrent neural network, trained on simulated data, to predict the indoor temperatures. This model is then used to train two DRL agents, with or without expert guidance, for the optimal control of the supply water temperature. Our tests in a multi-apartment setting show that both agents can ensure a higher thermal comfort and at the same time a smaller energy cost, compared to an optimized baseline strategy.

1 Introduction

Space heating and cooling in buildings is well-known for representing a significant part of global CO_2 emissions. For instance, district heating alone in China consumes more energy than the entire United Kingdom [1]. As of today, operating such industrial heating networks is prone to energy losses due to complex nonlinear building thermal behaviours. Hence significant efficiency gains are possible resulting in a clear line of work to fight climate change: develop and implement advanced control strategies for an optimal operation. This is all the more relevant in China where 87% of heat production is either from coal or oil [2].

In a typical Chinese district heating system, heat is produced in a central location and conveyed towards substations and onwards for distribution to customers via a heat exchanger and a network of insulated pipes. The distribution network is organized in a feed-and-return line and contains two parts, the primary and secondary networks as shown in Figure 1. The secondary network contains temperature sensors T_s, for the fluid exiting the substation towards the feed line, and T_r for the fluid entering the heat exchanger back from the return line. As is the norm in China and unlike some contributions on optimal district heating such as [3–5], we assume that the network operator buys heat at a constant price and from a unique third-party producer. We focus thus on the control of the secondary indoor temperatures: heat must be delivered to each apartment on the feed line to ensure that every indoor temperature is within an admissible range. In particular, any temperature above the upper bound results in both a waste of energy and an economic loss for the utility - most often, heating fees in China are a flat cost per square meter, regardless of the actual energy consumption. However, it is not possible to individually control the thermal behaviour of every apartment. Instead, an operator should control the indoor temperatures with two commands located inside the substation: (i) the supply temperature T_s, by acting on a control valve on the primary side, and (ii) the flow rate of the fluid, thanks to a pump at the inlet bound of the secondary side of the heat exchanger.

The state-of-the-art industrial control strategies rely on a relationship, called the water curve, which is tuned by an expert, between T_s and the outdoor temperature T_o. This paper investigates how to improve this strategy to ensure both higher thermal comfort and smaller energy and operation cost.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

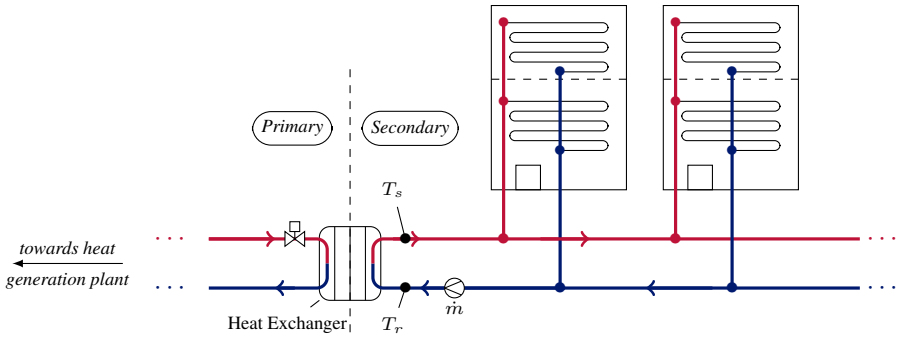

Figure 1: A district heating network.

2 Approach

2.1 Control strategy

We apply a Reinforcement Learning (RL) paradigm [6], where an *agent* learns a control strategy (*policy*) by interacting with the environment - here, the set of rooms heated by the network. The problem is modelled as a Markov Decision Process: the agent receives an observation of the *state* of the environment, chooses an *action* and receives as a result a *reward* from the environment. The best control strategy maximizes the expected cumulative discounted reward over the lifetime of the agent. Learning such a policy requires first to derive a model of the environment, to predict the indoor temperatures from the commands and the weather conditions. This model is described in Section 2.2.

At time t, the state is a vector s_t containing the outdoor temperature $T_{o,t}$, supply water temperature $T_{s,t}$, time of the day and indoor temperatures $T_{in,t}^{(j)}$ for every room $j \in \{1, \dots, N\}$ in the network. s_t contains both present and past n measurements of these quantities. At an hourly time step, a history of 24 hours is used to form s_t. At that same hourly time step, the agent is asked to select an action a_t. The flow rate being kept constant, the action is restricted to the supply temperature $T_{s,t}$. Two discrete action spaces, with $T_s(^\circ\mathrm{C}) \in \{20, 21, \dots, 50\}$, are considered. Agent 1 is the standard strategy while Agent 2 is a finetuning of the baseline control strategy (cf Section 2.3):

1. Agent 1: to enforce the smoothness of the control signal, the action is limited to the increments $a_t = T_{s,t} - T_{s,t-1}$ where $a_t \in \mathcal{A} := \{0, \pm 0.5, \pm 1, \pm 1.5, \dots, \pm 3\}$.
2. Agent 2: the discrete action is the difference $a_t = T_{s,t} - T_{s,t}^b$ where $a_t \in \mathcal{A}$ and T_s^b is the estimated baseline supply temperature.

Finally, the agent selects the action in order to maximize the expected cumulative discounted reward function $R = \sum_{t=0}^{T} \gamma^t r(s_t, a_t)$ over T time steps in the heating season. In the sequel, the discount factor is set to $\gamma = 0.9$, which corresponds to an agent that adapts its behaviour to the expected reward for the next 30 hours. The reward r penalizes deviations from a target temperature $\mathcal{T}$:

$$r(a_t, s_t) = -\sum_{j=1}^{N} |T_{in,t}^{(j)} - \mathcal{T}_t^{(j)}|. \tag{1}$$

We use Deep Reinforcement Learning (DRL), to train the different agents. DRL has proven to be a successful algorithm in various domains such as games, robotics or demand response [7]. In particular, we train Deep Q-Networks (DQNs, [8, 9]). For each training episode, a weather file is randomly chosen from a set of 7 cities in China to avoid overfitting the local climate, and an entire heating season is simulated. The weather measurements for testing the agents come from an eighth city, Yuncheng. Some statistics summarizing the climate in these cities are gathered in Appendix A.

2.2 Model identification

Consider a six-story building with three apartments per level - facing either the Eastern, Southern or Western direction. Heat is provided from a district heating substation and supplied to the apartments

through a high-inertia radiant heating system. The detailed building dynamics are simulated from an in-house model developed using Dymola, a commercial physics-based modelling and simulation tool, and calibrated against real operation data to represent a variety of indoor temperatures found in a district rather than in a single building. Seven apartments are thus left empty, to represent thermal losses and eleven indoor temperatures are simulated at an hourly time-step from two sets of inputs: the weather conditions (hour of the day, outdoor temperature T_o, solar flux and angles) and the commands (supply water temperature T_s and mass flow rate $\dot{m}$). Another output of the model is the water return temperature T_r. The model simulates a heating season from mid-December to mid-March for a total of 83 days (2002 hours).

The detailed physical model is then used as a data generator process to train a statistical model predicting the indoor and return temperatures. This model is a recurrent neural network (RNN, [10]) with two layers of 32 Long-Short Term Memory units [11], a class of neural network able to catch long-term dependencies in sequential data. Due to the high inertia of the system, the inputs are the same as for the detailed model, except that a sequence of 120 time-steps ($t-119$ to t) is used to make predictions at time t. 1,373 series are generated, of which 80% are for training, 10% for validation and 10% for testing. Each simulation has a random command and a weather file chosen from a random location in China (see Appendix A). Implemented in Python with `keras-tensorflow` [12], the model achieves a mean absolute test error of $0.110^{\circ}\mathrm{C}$ with a standard deviation of $0.132^{\circ}\mathrm{C}$.

Switching from expert model to RNN reduces the prediction time - about 30 times faster on CPU - which enables an efficient training of the agent. Although a challenge in practice, RNN networks can also be fine-tuned to on-site measurements with transfer learning, ensuring thus a higher scalability.

2.3 Baseline model and metrics

The baseline model is a linear water curve $T_s^b = \alpha + \beta T_o$, α, β being found by minimizing $\sum_t -r_t$, r_t as in (1), with Particle Swarm Optimization [13]. The optimal linear water curve is an advanced industry strategy, also called reactive control [5]. Besides, we implement and tune a PID controller, a frequently used strategy for temperature control [14]. All computations are carried out in Python.

The different policies are compared in terms of (i) thermal comfort, i.e. deviation from target temperature and stability of indoor temperatures, (ii) energy cost and (iii) estimated CO_2 cost. The energy cost is computed from the temperature difference $T_s - T_r$ with T_r predicted by the RNN model. The CO_2 cost is the amount of CO_2 emissions for one square meter of heated surface and per heating season, under the standard industry assumption that it requires 80 $\mathrm{kWh}/\mathrm{m}^2$ of heat from a coal-fired co-generation power plant, a wide-spread technology in China. Detailed computations are presented in Appendix A.

3 Results and Discussion

We consider first the control of a single apartment with constant target $\mathcal{T} \equiv 18^{\circ}\mathrm{C}$. It can be seen from Figure 2 that Agent 1 uses the extra degree of freedom on T_s to maintain a more stable temperature, closer to the target, hence a higher thermal comfort. Besides, the energy cost of Agent 1 is 2.73% smaller than the baseline. Similar conclusions apply to Agent 2 (see Appendix A, Figure 4).

Next, Table 1 summarizes the metrics for controlling the 11 indoor temperatures at the same time, with $\mathcal{T} \equiv 18^{\circ}\mathrm{C}$. The Agents achieve a better control of indoor temperatures, combined with energy and CO_2 savings: controlling a heating network serving one million square meters (about 20,000 customers) with Agent 2 would save about 495 tons of CO_2 per season. Using the expert baseline strategy to guide the actions of the agent is also beneficial, as illustrated by the improved performances of Agent 2. This might help increasing the acceptability of reinforcement learning for the network operator, since in this case the action is always at most $3^{\circ}\mathrm{C}$ different from the expert choice. Finally, the PID controller performs slightly better than the baseline, but not as well as the Agents.

It can be noted that the environment contains little flexibility - a unique heat source at fixed price, no storage units - whereas advanced controls are designed to unlock it [3, 15]. Hence, robust heuristics like the optimal water curve are here difficult to improve further. Yet, flexibility can be achieved by modifying the agent's reward to allow more refined dynamic control targets, e.g. with a two-level target temperature $\mathcal{T} = 17^{\circ}\mathrm{C}$ (night) and $\mathcal{T} = 18^{\circ}\mathrm{C}$ (day), Agent 1 achieves 6.6% energy savings.

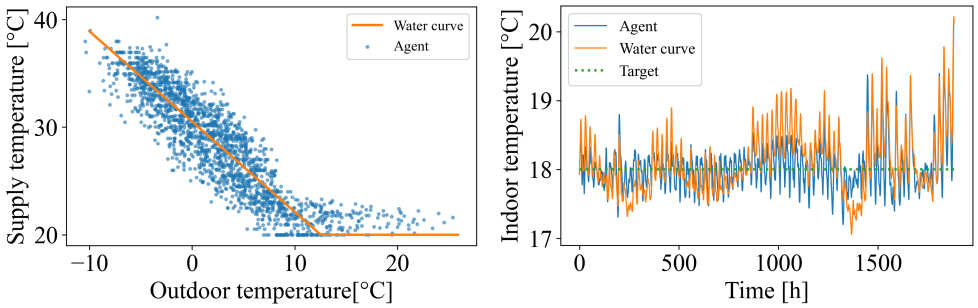

Figure 2: Agent 1 and Baseline performances for one apartment. (left) Control strategies T_s vs T_o and (right) trajectories of T_{in}. Constant target at 18°C. Best viewed in colors.

The good performances of the agents are in line with several other recently published papers applying reinforcement learning to controlling an indoor temperature, with smaller energy gains due to the lack of flexibility of the environment and an optimized baseline. Besides, our approach differs from these contributions in some aspects. First of all, whereas most studies focus on the thermal behaviour of a unique building, e.g. [16–19], our system is a whole substation, with the 11 apartments originally tuned to represent the variety found in a district. District heating system are studied in [3–5, 20] with promising results in terms of control, but with no analysis of the effect of their policies on indoor temperatures. Secondly, these references assume a simple rule-based strategy for baseline, or a manual control [14]. For district heating however, using water curves for setting the supply temperature is common practice, with parameters tuned by hand from a set of a dozen indoor temperature sensors representative of the district thermal behaviour. By optimizing these parameters, our baseline strategy is thus an improvement compared to the industry standards, in order to better assess the potential gain due to reinforcement learning.

Another key feature of applying reinforcement learning to real-world problems is the design of the reward function. Most references build the reward as an explicit trade-off between energy cost and thermal comfort with careful weighting of the two contributions, see e.g. [14, 16–18]. On the contrary, our approach was to define a reward that depends solely on the target temperature specified by the contract between utility and customers, and to evaluate whether the agents can lower the energy cost as a side effect. Indeed, when adding an energy cost in this low flexibility environment, the Agents maintain a mean indoor temperature constant at the lowest possible level; this effect can also be achieved by the baseline by lowering the target temperature. Moreover, we find in our experiments that the reward function (1) is more stable and robust to different weather conditions, while still maintaining an advantage in terms of both energy cost and thermal comfort.

Nevertheless, our results suggest that deep reinforcement learning, by understanding the dynamics of the system, is a suitable tool for controlling district heating networks, maintaining thermal comfort while reducing energy cost. In order to be applied to an actual network, the first step is to deploy onsite outdoor and indoor temperature sensors. Next, either the RNN model or a lightweight statistical model (e.g. equivalent RC electrical networks) is finetuned on the operation data. Based e.g. on a cloud infrastructure to store the measurements, the agents, whether they are DQN or more recent agents such as DDPG [21], can finally be deployed for controlling the substation [20].

Table 1: Performances of the control strategies in the multi-apartment setting, for $\mathcal{T} \equiv 18°C$. MAE: mean absolute error, *std*: standard deviation. Best performance is emphasized in **bold**.

	MAE T_{in} (°C)	**std** T_{in} (°C)	**Energy gain** (%)	**CO_2 saved** (g/m^2)
Baseline	0.599	0.755	0	0
PID	0.584	0.742	0.95	215
Agent 1	0.549	0.699	2.15	486
Agent 2	**0.545**	**0.692**	**2.19**	**495**

References

[1] *District Energy Systems in China*, IEA, Paris, 2017. [Online]. Available: https://www.iea.org/reports/district-energy-systems-in-china

[2] IEA, 2018. [Online]. Available: https://www.iea.org/data-and-statistics?country=CHINAREG&fuel=Electricity%20and%20heat&indicator=Heat%20generation%20by%20source

[3] L. Giraud, M. Merabet, R. Baviere, and M. Vallée, "Optimal control of district heating systems using dynamic simulation and mixed integer linear programming," in *Proceedings of the 12th International Modelica Conference, Prague, Czech Republic, May 15-17, 2017*, no. 132. Linköping University Electronic Press, 2017, pp. 141–150.

[4] S. J. Cox, D. Kim, H. Cho, and P. Mago, "Real time optimal control of district cooling system with thermal energy storage using neural networks," *Applied energy*, vol. 238, pp. 466–480, 2019.

[5] E. Saloux and J. Candanedo, "Optimal rule-based control for the management of thermal energy storage in a canadian solar district heating system," *Solar Energy*, vol. 207, pp. 1191–1201, 2020.

[6] R. S. Sutton and A. G. Barto, *Introduction to reinforcement learning*. MIT press Cambridge, 1998, vol. 135.

[7] Y. Li, "Deep reinforcement learning: An overview," *arXiv preprint arXiv:1701.07274*, 2017.

[8] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

[9] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[10] R. J. Williams and D. Zipser, "A learning algorithm for continually running fully recurrent neural networks," *Neural computation*, vol. 1, no. 2, pp. 270–280, 1989.

[11] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[12] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

[13] R. Eberhart and J. Kennedy, "A new optimizer using particle swarm theory," in *Proc. 6th Int. Symp. Micro Mach. Human Sci. (MHS95)*. IEEE, 1995, pp. 39–43.

[14] Z. Zhang, A. Chong, Y. Pan, C. Zhang, and K. P. Lam, "Whole building energy model for hvac optimal control: A practical framework based on deep reinforcement learning," *Energy and Buildings*, vol. 199, pp. 472–490, 2019.

[15] A. Vandermeulen, B. van der Heijde, and L. Helsen, "Controlling district heating and cooling networks to unlock flexibility: A review," *Energy*, vol. 151, pp. 103–115, 2018.

[16] T. Wei, Y. Wang, and Q. Zhu, "Deep reinforcement learning for building hvac control," in *Proceedings of the 54th Annual Design Automation Conference 2017*, 2017, pp. 1–6.

[17] A. Nagy, H. Kazmi, F. Cheaib, and J. Driesen, "Deep reinforcement learning for optimal control of space heating," *arXiv preprint arXiv:1805.03777*, 2018.

[18] Z. Zhang, A. Chong, Y. Pan, C. Zhang, S. Lu, and K. P. Lam, "A deep reinforcement learning approach to using whole building energy model for hvac optimal control," in *2018 Building Performance Analysis Conference and SimBuild*, 2018.

[19] R. Jia, M. Jin, K. Sun, T. Hong, and C. Spanos, "Advanced building control via deep reinforcement learning," *Energy Procedia*, vol. 158, pp. 6158–6163, 2019.

[20] T. Zhang, J. Luo, P. Chen, and J. Liu, "Flow rate control in smart district heating systems using deep reinforcement learning," *arXiv preprint arXiv:1912.05313*, 2019.

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

[22] Y. Zhang, Q. Lyu, Y. Li, N. Zhang, L. Zheng, H. Gong, and H. Sun, "Research on down-regulation cost of flexible combined heat power plants participating in real-time deep down-regulation market," *Energies*, vol. 13, no. 4, p. 787, 2020.

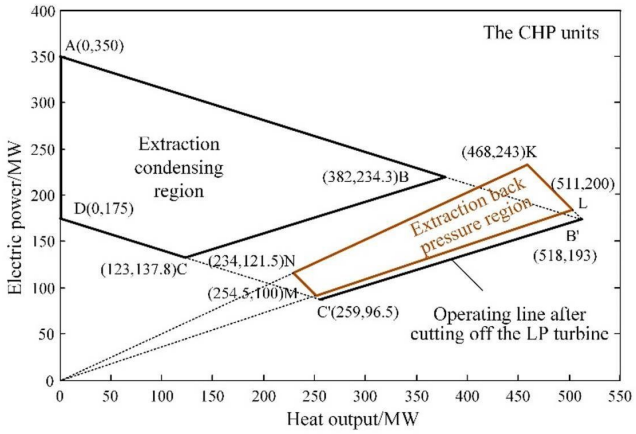

Figure 3: Feasible operating regions for Combined Heat and Power Units. Source: [22].

A Supplementary material

A.1 Weather information

The weather files used in our numerical experiments come from eight cities in China. Yuncheng's climate is used only for testing the different control strategies, while measurements from the seven other cities are used for training only. Table 2 summarizes their climate in terms of indoor temperature and solar radiation.

Table 2: Measured weather conditions averaged over one heating season. T_o (°C): outdoor temperature, $\max \phi$ (W.h/m^2): daily maximum global horizontal solar flux. *std:* standard deviation.

	Beijing	**Chengdu**	**Harbin**	**Shenyang**	**Shijiazhuang**	**Xian**	**Yuncheng**	**Zhengzhou**
mean T_o	6.92	12.36	-3.37	1.87	8.27	8.40	2.53	8. 81
(std T_o)	(7.58)	(4.78)	(10.01)	(1.87)	(7.20)	(6.67)	(5.81)	(6.69)
max ϕ	634	374	533	582	555	510	631	580
(std max ϕ)	(186)	(261)	(167)	(198)	(211)	(226)	(315)	(230)

A.2 Metrics computation

The energy cost is computed as follows. With T_s (respectively T_r) denoting the secondary supply (respectively return) water temperature, the heat transferred from the primary to the secondary side through the heat exchanger is:

$$Q = \dot{m} \cdot c_p \cdot (T_s - T_r),$$

where $\dot{m}$ is the water flow rate in kg/s, $c_p = 4180$ J/kg/K and Q is the heat duty (W).

To estimate the CO_2 cost, we assume that heat is produced by a coal-fired co-generation power plant. Under this technology, we make the assumption, standard in industry, that the primary heat consumption is 80 kWh/m^2. If a control policy saves $p\%$ energy, this corresponds thus to $80 \cdot p$kWh/m^2 of saved primary heat. Using the feasible operating region of combined heat and power plants shown in Figure 3, reducing the heat consumption by $80 \cdot p$ increases the electricity generation by $80 \cdot p \cdot \Delta$, where $\Delta = \left|\frac{350-234}{0-382}\right|$. Finally, we assume than producing one more kilowatt-hour of electricity without increasing the coal consumption saves 930 g of CO_2. Hence the estimation of total amount of saved CO_2 emissions:

$$930 \times \left|\frac{350-234}{0-382}\right| \times 80 \times p \quad (\text{gCO}_2/\text{m}^2)$$

A.3 Additional figures and DQN hyperparameters

The DQN has a learning rate set to 0.001, buffer size to 1000000, batch size to 32, 300 training episodes (1883 time steps per episode, after initialization during 119 steps); the initial (respectively final) value of random action probability is 1.0 (resp. 0.1), fraction of entire training period over which the exploration rate is annealed is 0.8, the target network is updated every 200 steps. The Q-network has two layers of 64 neurons each.

The results for one apartment controlled by Agent 2 are displayed in Figure 4.

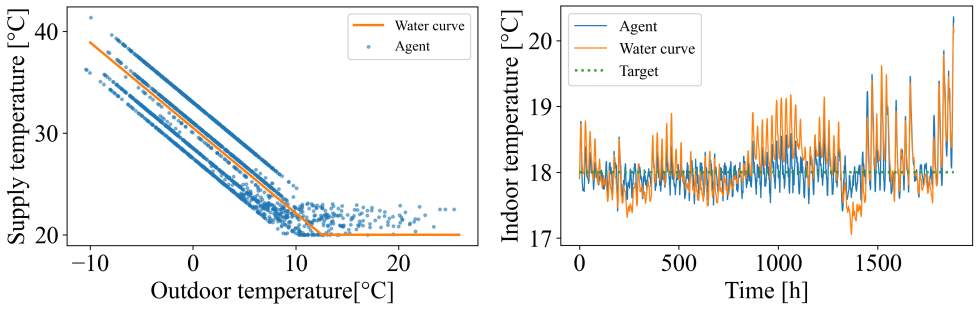

Figure 4: Results for a single apartment, from Agent 2 and baseline strategy. (left) Control strategy T_s vs T_o and (right) performance in terms of T_{in}, through the heating season. Best viewed in colors.

Additional figures are provided for the experiments with multiple apartments: Figure 5 shows the reward of both agents during the 300 episodes of training and Figure 6 shows the optimal control strategies and indoor temperatures for the two agents and the baseline.

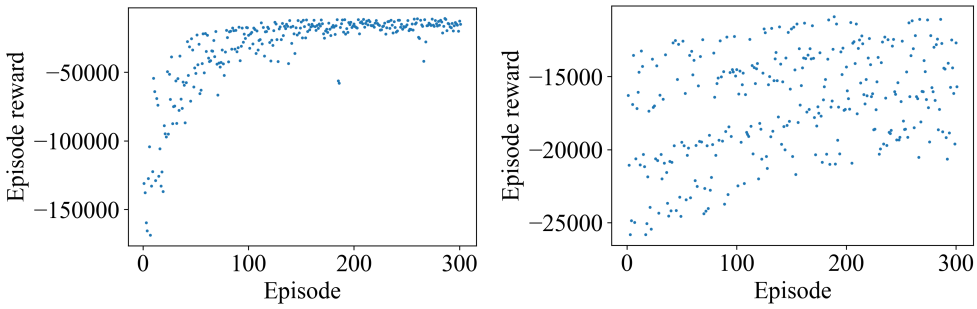

Figure 5: Rewards during training of Agents 1 (left) and 2 (right) for controlling all apartments.

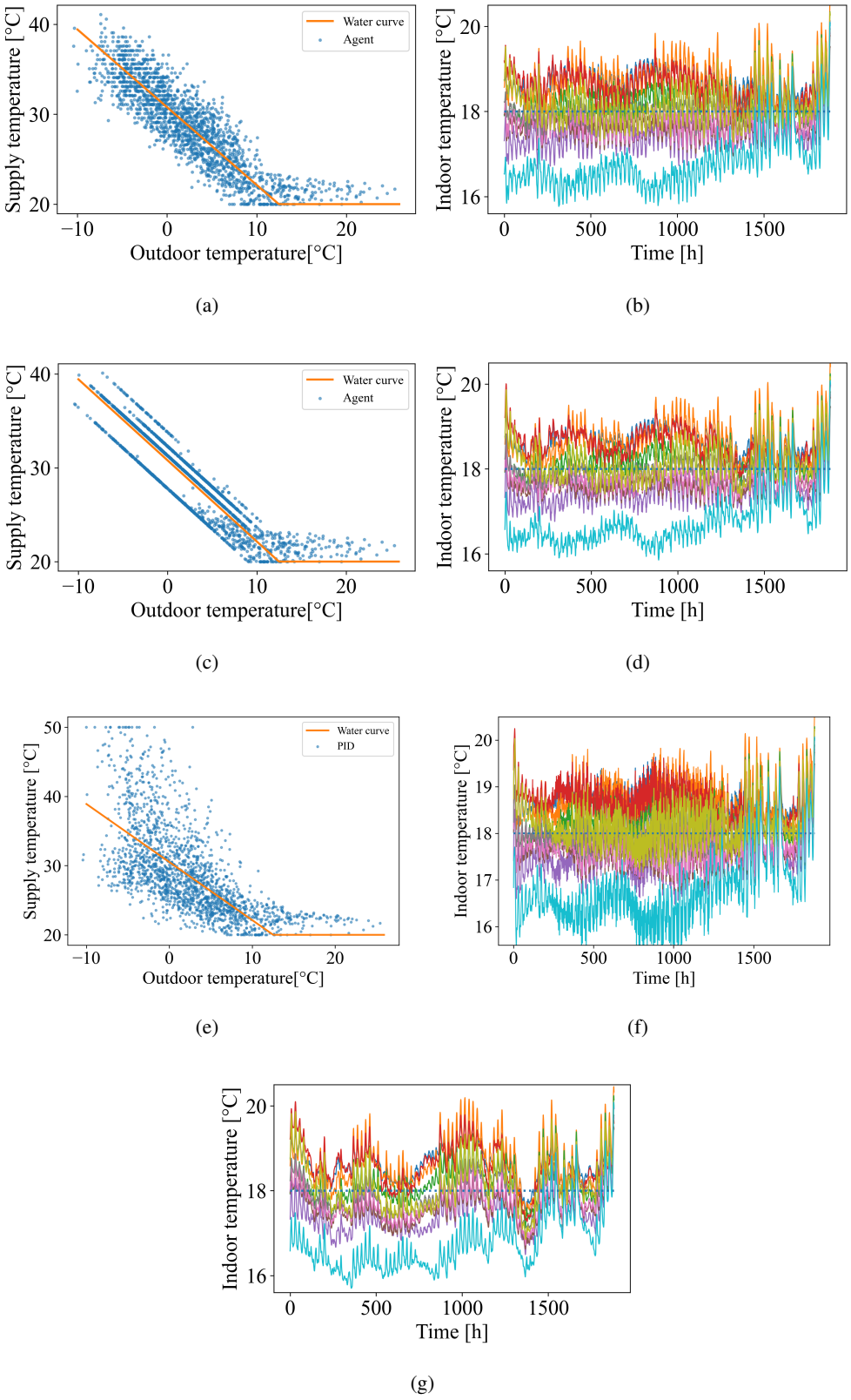

Figure 6: Control strategies of (a) Agent 1 (c) Agent 2 and (e) PID against the water curve, for multiple apartment and $\mathcal{T} \equiv 18°\text{C}$. Performances of (b) Agent 1, (d) Agent 2 (f) PID and (g) the baseline in terms of T_{in} for multiple apartments during the heating season. Best viewed in colors.