*energies*

MDPI

*Article*

# Building Heat Demand Prediction Based on Reinforcement Learning for Thermal Comfort Management

**Chendong Wang [1], Lihong Zheng [1,2], Jianjuan Yuan [3], Ke Huang [1,4] and Zhihua Zhou [1,*]**

[1] Tianjin Key Laboratory of Indoor Air Environmental Quality Control, Key Laboratory of Efficient Utilization of Low and Medium Grade Energy, School of Environmental Science and Engineering, Tianjin University, Tianjin 300350, China
[2] Tianjin Eco-City Green Building Research Institute, Tianjin 300467, China
[3] School of Energy and Environmental Engineering, Hebei University of Technology, Tianjin 300401, China
[4] Capital Construction Department, Tianjin University of Technology, Tianjin 300384, China
* Correspondence: zhuazhou@tju.edu.cn

**Abstract:** The accurate prediction of building heat demand plays the critical role in refined management of heating, which is the basis for on-demand heating operation. This paper proposed a prediction model framework for building heat demand based on reinforcement learning. The environment, reward function and agent of the model were established, and experiments were carried out to verify the effectiveness and advancement of the model. Through the building heat demand prediction, the model proposed in this study can dynamically control the indoor temperature within the acceptable interval (19–23 °C). Moreover, the experimental results showed that after the model reached the primary, intermediate and advanced targets in training, the proportion of time that the indoor temperature can be controlled within the target interval (20.5–21.5 °C) was over 35%, 55% and 70%, respectively. In addition to maintaining indoor temperature, the model proposed in this study also achieved on-demand heating operation. The model achieving the advanced target, which had the best indoor temperature control performance, only had a supply–demand error of 4.56%.

**Keywords:** reinforcement learning; heat demand prediction; on-demand heating operation; deep learning

check for updates

**Citation:** Wang, C.; Zheng, L.; Yuan, J.; Huang, K.; Zhou, Z. Building Heat Demand Prediction Based on Reinforcement Learning for Thermal Comfort Management. *Energies* **2022**, *15*, 7856. https://doi.org/10.3390/en15217856

Academic Editors: Xiangfei Kong and Tailu Li

Received: 14 September 2022
Accepted: 17 October 2022
Published: 23 October 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

## 1. Introduction

### 1.1. Backgrounds

In recent years, due to rapid industrialization, urbanization and population expansion, the energy demand has significantly increased globally. China has put forward the "3060 target" in the "14th Five-Year Plan" and has listed it as one of the key tasks, striving to reach carbon emission peak by 2030, and achieving carbon neutrality by 2060 [1]. In China, building operation energy consumption accounts for 30% of the total energy consumption, hence the realization of "3060 target" is inseparable from the concerted efforts of building energy conservation [2]. China has become one of the largest heating markets in the world. In 2019, the heating area in northern China was 15.2 billion $m^2$, the heating energy consumption was 213 million tce, accounting for 20.9% of the total building energy consumption in China, and the carbon emission intensity was 36 kg $CO_2/m^2$ [3]. Therefore, facing the huge pressure of resources, environment, and indoor thermal comfort, the urban heating industry in China are actively exploring clean heating methods such as geothermal energy, solar energy, industrial waste heat heating, etc. At the same time, it is also seeking energy-saving control strategies to achieve refined heating management and on-demand heating operation, supplying heat according to the actual demand of heat users.

Although the heating energy consumption in China is high, the building thermal comfort in China is still lower than that of developed countries and some Asian countries [4,5]. Research on indoor temperature distribution showed that the indoor temperature was often higher than 22 °C or lower than 18 °C in northern China, accounting for 50–70% time

of the heating season. It indicated that most of the heating systems in China have not achieved on-demand heating operation, and there was huge energy saving potential. With the development of information technology such as machine learning, big data, artificial intelligence, etc., refined management has become a novel concept and new mode for heating. The accurate prediction of heating heat demand plays the critical role in refined management of heating, which is the basis for on-demand heating operation.

In the past few years, many studies have carried out extensive research on building heat demand prediction through black-box models. A variety of black-box model algorithms have been investigated, including statistical methods, e.g., multiple linear regression (MLR) [6], autoregression integrated moving average (ARIMA) [7]; traditional machine learning algorithms, e.g., support vector regression (SVR) [8]; ensemble methods, e.g., random forest [9], extreme gradient boosting [10]; and neural-network-based methods, e.g., artificial neural network (ANN) [11], convolutional neural networks (CNN) [12], long short-term memory (LSTM) [13–15], etc. Most of these research studies focused on improving prediction accuracy through algorithm optimization or novel advanced algorithms. The heat demand pattern of the building can be learned by model training using historical data, and hence the future heat demand can be predicted. This method is based on the consistency between the training model and the historical operating conditions. However, the evaluation of the operation performance of the historical operating conditions is ignored in this process, and the model achieved by this method cannot realize the transition from high-energy consumption operation to low-energy consumption operation. In other words, if the energy consumption of the historical operating conditions is high, the energy consumption guiding by the model also will be high during practical operation. Therefore, even if the trained model fits the historical data well and the prediction accuracy is high, it does not mean that the model can be used to achieve on-demand heating operation. It is still urgent to develop a model framework that can predict the heat demand in real time according to the target indoor thermal comfort performance.

### *1.2. Reinforcement Learning*

In recent years, reinforcement learning (RL) has become one of the important developments in the field of machine learning. RL achieves autonomous learning through continuous interaction between the agent and the environment and continuous strategy optimization. It is a trial-and-error mechanism that interacts with the environment and maximizes cumulative rewards. The combination of RL and deep learning, i.e., deep reinforcement learning (DRL), has shown superior learning ability and accuracy [16–18], owing to its strong nonlinear fitting ability of deep neural networks and excellent decision-making ability of RL. DRL has been widely used and made impressive achievements in automatic control such as mountain climbing [19], human–computer behavior interaction [20] and game process [21], and automatic production and control in the industrial field [22,23].

It is evident from the above application scenarios that RL and DRL are good at solving sequence optimization problems. However, it is different from traditional sequential optimization methods that the agent in DRL seeks the highest cumulative reward and is therefore able to consider multi-step situations in the future and make optimal decisions. Therefore, more and more scholars have begun to apply DRL to load prediction and operation optimization problems of building energy system. Liu et al. [24] proposed three DRL algorithms for single-step and multi-step prediction of building energy consumption, respectively. The results showed that the proposed algorithm was more accurate than the supervised learning algorithm. Mocanu et al. [25] investigated the performance of different RL and DRL methods on the operational optimization of building energy systems. The results showed that the deterministic policy gradient (DPG) algorithm was more stable than the deep Q-network (DQN) algorithm and DPG algorithm reduced building operating costs by 27.4% and peak loads by 26.3%. Mason et al. [26] reviewed the application of RL in building energy management systems, such as in HVAC, photovoltaic, power grids, and lighting. Baghaee et al. [27] used a Q-learning-based model to regulate indoor $CO_2$

concentration and the HVAC system simultaneously. Faddel et al. [28] proposed a Q-learning method to minimize building energy consumption while ensuring the thermal comfort of occupants. It is apparent from the above literature research that few research studies investigated building heat demand prediction, and current research on combining RL with building heat demand prediction is still in its infancy.

### *1.3. Research Objectives*

This paper proposed a prediction model framework based on RL and further evaluate the performance to achieve on-demand heating operation. The contributions of this study can be summarized as follows:

- Based on the DRL algorithm, this study addressed accurate prediction of indoor heat demand to achieve on-demand heating operation.
- The performance of the DRL model was analyzed, including model training, indoor temperature control and the supply–demand error.
- Appropriate reward function was proposed for the application of the DRL model in building heat demand prediction tasks.
- A novel evaluation criterion was applied for the prediction model, the indoor thermal comfort performance (indoor temperature) was used instead of prediction accuracy to evaluate model performance.

The structure of this paper is as follows. Section 2 presents the methodology of this study, including the theoretical background of RL, DRL, deep deterministic policy gradient (DDPG) and twin delayed deep deterministic policy gradient (TD3), the architecture of proposed prediction model, experiment schemes, model settings and evaluation indices. Then, Section 3 describes the experiment results and evaluates the performance of the elaborate experiments. Finally, all findings and implications are summarized in Section 4.

## 2. Methods

### *2.1. Reinforcement Learning*

RL is a branch of machine learning, which is mainly composed of reward, agent and environment, as shown in Figure 1. The agent picks up new information through trial and error, then directs its behavior to obtain maximum cumulative reward based on the interaction with the environment. RL problem usually can be represented by a Markov decision process (MDP), represented by a quadruple: $M = \langle S,\ A,\ R,\ P \rangle$, where $S$ represents the state space; $A$ represents the action space; $R$ represents the immediate reward received after transitioning from state $S_t$ to state $S_{t+1}$; and $P$ represents the probability that action $a$ in state $S_t$ will lead to state $S_{t+1}$.

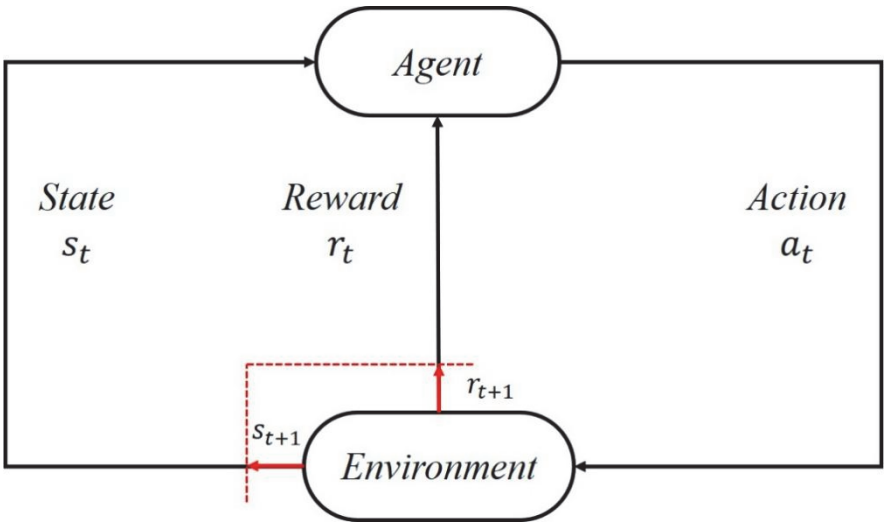



**Figure 1.** The framework of reinforcement learning.

The state value function $V^{\pi}(s)$ and action value function $Q^{\pi}(s,a)$ are two closely related value functions that are used to create a typical RL algorithm, which are described in Equations (1) and (2) [29], where $\pi$ stands for a specific policy function; $\gamma$ represents the discount factor which specifies the weighting relationship between future rewards and immediate reward, $\gamma \in [0,1)$. Equations (3) and (4) [29] are the Bellman equations for state value and action value, respectively, which extract the iterative connection of the value function between two adjacent time steps.

$$V^{\pi}(s) = E_{\pi}\left\{\sum_{k=0}^{\infty} \gamma^{k} r_{t+k+1} \mid s_t = s\right\} \tag{1}$$

$$Q^{\pi}(s,a) = E_{\pi}\left\{\sum_{k=0}^{\infty} \gamma^{k} r_{t+k+1} \mid s_t = s, a_t = a\right\} \tag{2}$$

$$V^{\pi}(s) = E_{\pi}\left\{r_{t+1} + \gamma V^{\pi}(s') \mid s_t = s, s_{t+1} = s'\right\} \tag{3}$$

$$Q^{\pi}(s,a) = E_{\pi}\left\{r_{t+1} + \gamma Q^{\pi}(s') \mid s_t = s, a_t = a, s_{t+1} = s'\right\} \tag{4}$$

### *2.2. Deep Deterministic Policy Gradient*

Traditional RL methods, such as DQN, usually rely on tables to process a finite number of actions and states, hence they are unable to handle continuous actions or states, which limits their further application and development. The DDPG algorithm is a DRL method which has the same theoretical foundation with RL methods, but instead of using a table to estimate cumulative rewards, it introduces a deep neural network, which enables it to expand to a continuous action space [30].

The DDPG algorithm is based on the actor–critic method and employs a DPG to improve the DQN algorithm. The agent includes two kinds of neural networks, Actor and Critic. The Actor is responsible for selecting actions and it outputs a continuous action vector; while the Critic is responsible for estimating the action value function $Q^{\pi}(s,a)$ of each action. The Critic updates according to Equation (5) [30], which is called the temporal-difference (TD) algorithm [29]. The update target of the Actor depends on the calculation results of the Critic.

$$\delta_t = r_{t+1} + \gamma V(s_{t+1}) - V(s_t) \tag{5}$$

The balance of exploration and utilization of RL is a crucial but difficult task [31], because whether the agent has thoroughly explored the environment influences the ultimate performance of the model. Insufficient exploration may prevent the agent from obtaining valid information, while excessive exploration can significantly influence the convergence speed. Due to the action space is continuous in the DDPG algorithm, the action with the maximum value function cannot be directly selected from the action space, and hence the greedy strategy used in the DQN algorithm can no longer be used. Therefore, random noise values are added to the output of the Actor for exploration in the DDPG algorithm, as given in Equation (6) [30].

$$\mu'(s_t) = \mu(s_t \mid \theta_t^{\mu}) + \mathcal{N} \tag{6}$$

where, $\mathcal{N}$ can be any type of noise based on the situation of the environment, such as Gaussian noise [32].

In addition, the DDPG algorithm introduces double Critic networks to optimize the performance. If the agent only contains single Critic, the training process can be very unstable. Because the Critic network is still used to calculate its target value (TD error) for updating. Therefore, a target network is established to calculate the TD error specially. The target network has the same structure and initial parameters as those of the original Critic network, but it is not updated with the gradient of the original network [21]. The target network uses the delay update strategy, it copies the parameters of the original network gradually to implement the update.

### 2.3. Twin Delayed Deep Deterministic Policy Gradient

For DRL problems, the calculation process has problems with overestimation and variance [33]. Overestimation often leads to the strategy updating in the wrong direction, whereas too much variance leads to instability in training and increases the convergence time. The TD3 algorithm sets up twin Critic networks with the same structure to calculate the TD error and selects the minimum value of the two Critic networks for the update, as given in Equation (7) [33].

$$y_1 = r + \gamma \min_{i=1,2} Q_{\theta_i'}(s_{i+1}, \mu'(s_{i+1}|\theta^{\mu'})) \quad (7)$$

In addition to using the twin Critic networks, the TD3 algorithm decreases the update frequency to make the update target more stable. Moreover, the TD3 algorithm does not update the other parameters until the current Critic is stable. Essentially, the TD3 algorithm is a DDPG algorithm with some minor changes, but it achieves very good results.

### 2.4. Proposed Building Heat Demand Prediction Mode

This study established a building heat demand prediction model framework based on RL, the environment, reward function and agent of the model were established. Proposed model can accurately predict the dynamic heat demand of indoor environment and realize on-demand heating operation.

#### 2.4.1. Environment

RL training needs to be achieved through the continuous interaction between the agent and the environment. Hence, this study built a simulation environment based on building physical parameters and a mathematical model of the heat transfer to cooperate with the agent for training.

Physical Parameters

(1) Building shape

The building simulated in this study is a simple residential building with only one floor and one room. Detailed parameters of the simulated building environment are given in Table 1.

(2) Parameters of building envelopes

**Table 1.** Detailed parameters of simulated building.

| Parameters | Values |
|---|---|
| Length | 4 m |
| Width | 3 m |
| Height | 2.8 m |
| Area of windows | 1.2 $m^2$ |

The parameters of each envelope are shown in Table 2 when setting up the simulated building environment.

**Table 2.** The parameters of each envelope.

| External Walls (from Outside to Inside) | | | | |
|---|---|---|---|---|
| **Materials** | **Thickness (mm)** | **Thermal Conductivity (W/(m·K))** | **Density ($kg/m^3$)** | **Heat Capacity (kJ/(kg·K))** |
| Cement mortar | 20 | 0.93 | 1800 | 1.05 |
| Reinforced concrete | 200 | 1.74 | 2500 | 0.92 |
| EPS insulation board | 80 | 0.039 | 30 | 1.38 |
| Cement mortar | 20 | 0.93 | 1800 | 1.05 |

**Table 2.** *Cont.*

| Roofs (from Top to Bottom) | | | | |
|---|---|---|---|---|
| **Materials** | **Thickness (mm)** | **Thermal Conductivity (W/(m·K))** | **Density (kg/m$^3$)** | **Heat Capacity (kJ/(kg·K))** |
| Cement mortar | 20 | 0.93 | 1800 | 1.05 |
| XPS insulation board | 75 | 0.03 | 30 | 1.38 |
| Reinforced concrete | 200 | 1.74 | 2500 | 0.92 |
| Cement mortar | 20 | 0.93 | 1800 | 1.05 |
| **Floors (from Top to Bottom)** | | | | |
| **Materials** | **Thickness (mm)** | **Thermal Conductivity (W/(m·K))** | **Density (kg/m$^3$)** | **Heat Capacity (kJ/(kg·K))** |
| Cement mortar | 20 | 0.93 | 1800 | 1.05 |
| EPS insulation board | 25 | 0.039 | 30 | 1.38 |
| Reinforced concrete | 150 | 1.74 | 2500 | 0.92 |
| **External Windows (5 mm/9 mm Argon/5 mm/9 mm Argon/5 mm)** | | | | |
| **Parameters** | **Heat Transfer Coefficient (W/(m$^2$ K)** | | **Solar Heat Gain Coefficient (−)** | **Effective Area Factor (−)** |
| **Value** | 1.0 | | 0.7 | 0.75 |

Mathematical Model

Building envelopes can be divided into high thermal storage building envelopes (such as external walls, roofs and floors) and low thermal storage building envelopes (such as glass). In this study, the heat transfer process of building envelopes with high thermal storage was assumed as unstable heat transfer, and the heat transfer process of building envelopes with low thermal storage is assumed to be steady-state heat transfer. The simulated object is a residential building, and the heat capacity particle system method is used to divide the building envelope structure and establish a mathematical model. The building envelope structure is split into several layers, and a temperature node is set in each layer, assuming that the heat capacity of the layer is concentrated on the temperature node. Therefore, the heat balance equation between the temperature nodes of each layer can be established according to the law of conservation of mass and the law of conservation of energy. In this study, the nodes of external walls, roofs and floors are divided as shown in Figures 2–4, respectively.

(1) Dynamic heat balance of external walls and roofs

Since the node division results of the external walls and roofs are similar, their heat balance transfer equations between the nodes are consistent, as given in Equations ((8)–(13)).

$$A_i h_{out}(T_{i_out} - T_{out}) + q_{solar,i} = A_i K_{i1-O}(T_{i_1} - T_{i_out}) \quad (8)$$

$$\rho_{i_1} c_{i_1} A_i \delta_{i_1} \frac{\partial T_{i_1}}{\partial t} = A_i K_{i2-1}(T_{i_2} - T_{i_1}) - A_i K_{i1-O}(T_{i_1} - T_{i_out}) \quad (9)$$

$$\rho_{i_2} c_{i_2} A_i \delta_{i_2} \frac{\partial T_{i_2}}{\partial t} = A_i K_{i3-2}(T_{i_3} - T_{i_2}) - A_i K_{i2-1}(T_{i_2} - T_{i_1}) \quad (10)$$

$$\rho_{i_3} c_{i_3} A_i \delta_{i_3} \frac{\partial T_{i_3}}{\partial t} = A_i K_{i4-3}(T_{i_4} - T_{i_3}) - A_i K_{i3-2}(T_{i_3} - T_{i_2}) \quad (11)$$

$$\rho_{i_4} c_{i_4} A_i \delta_{i_4} \frac{\partial T_{i_4}}{\partial t} = A_i K_{iin-4}(T_{i_in} - T_{i_4}) - A_i K_{i4-3}(T_{i_4} - T_{i_3}) \quad (12)$$

$$A_i h_{in}(T_{in} - T_{i_in}) = A_i K_{iin-4}(T_{i_in} - T_{i_4}) \quad (13)$$

where, $i$ is the subscript of each parameter, indicating external walls or roofs; $A_i$ is the area, m$^2$; $\delta_i$ is the thickness, m; $c_i$ is the heat capacity, J/(kg·K); $\rho_i$ is the density, kg/m$^3$; $T_i$ is the temperature of each node, °C; $T_o$ is the outdoor temperature, °C; $T_{in}$ is the indoor

temperature, °C; $T_{i_out}$ is the temperature of the outer surface immediately adjacent to the outdoor air, °C; $T_{i_in}$ is the temperature of the inner surface immediately adjacent to the indoor air, °C; $K_i$ is the conductive heat transfer coefficient of each layer, W/(m$^2$·°C), which is given in Equation (14); $q_{solar_i}$ is the heat absorbed by the outer surface from solar radiation, W, which is given in Equation (15); $h_{out}$ is the convective heat transfer coefficient with outdoor air, $h_{out}$ = 19 W/(m$^2$·°C); $h_{in}$ is the convective heat transfer coefficient with indoor air, $h_{in}$ = 8.7 W/(m$^2$·°C).

$$K = 1/\sum \frac{\delta_j}{\lambda_j} \tag{14}$$

where, $\lambda_j$ is the thermal conductivity of each layer $j$, W/(m·°C); $\delta_j$ is the thickness of each layer $j$, m.

$$q_{solar,i} = A_i \alpha_i I_{sun} \tag{15}$$

where, $\alpha_i$ is the insolation absorption coefficient; $I_{sun}$ is the solar intensity, W/m$^2$.

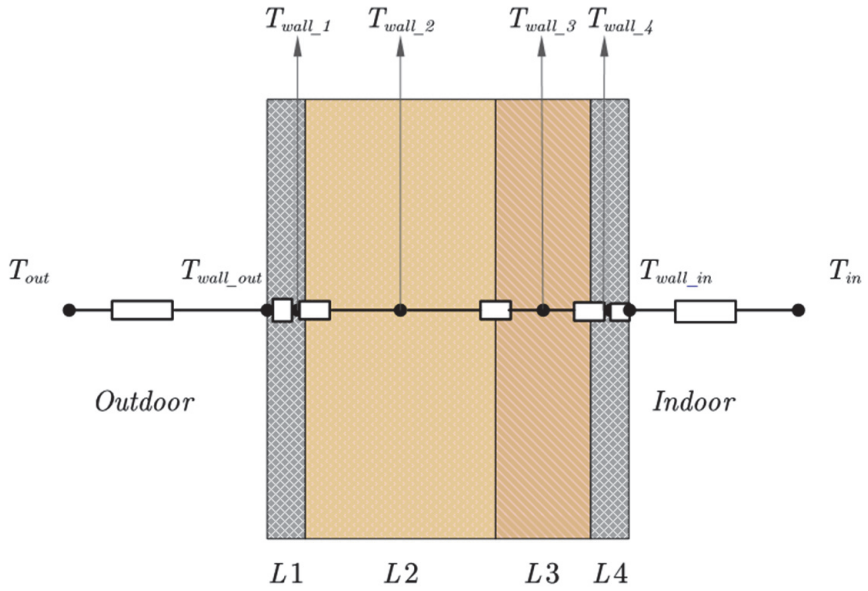



**Figure 2.** The nodes of external walls.

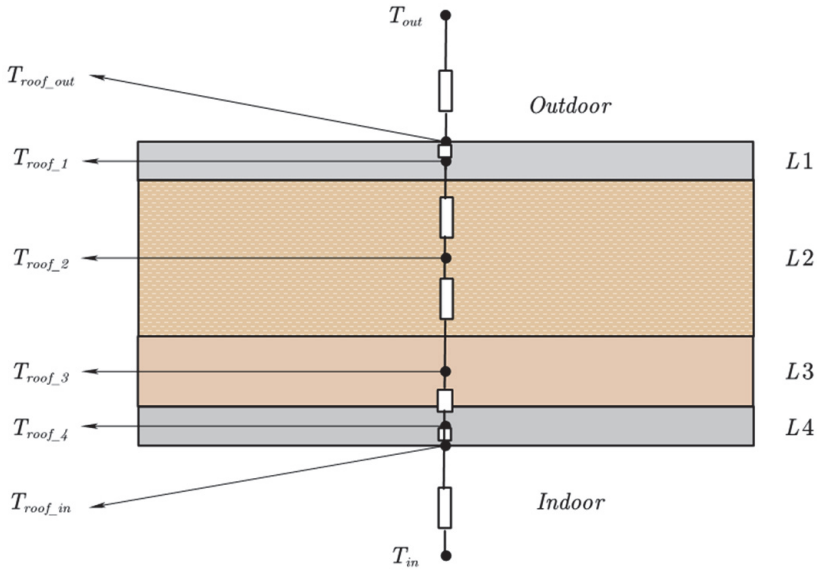



**Figure 3.** The nodes of roofs.

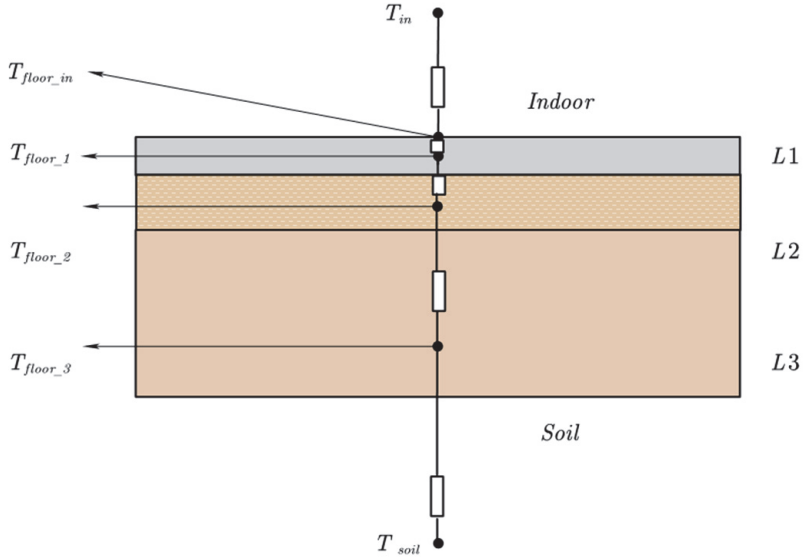



**Figure 4.** The nodes of floors.

(2) Dynamic thermal balance of floors

Due to the outermost layer of the floor contacts with the underground soil, the heat balance transfer equations between the nodes are different from external walls and roofs, as given in Equations (16)–(19).

$$A_f h_{in}\left(T_{in} - T_{f_in}\right) = A_f K_{fin-1}\left(T_{f_in} - T_{f1}\right) \tag{16}$$

$$\rho_{f_1} c_{f_1} A_f \delta_{f_1} \frac{\partial T_{f_1}}{\partial t} = A_f K_{fin-1}(T_{f_in} - T_{f_1}) - A_f K_{f1-2}(T_{f_1} - T_{f_2}) \tag{17}$$

$$\rho_{f_2} c_{f_2} A_f \delta_{f_2} \frac{\partial T_{f_2}}{\partial t} = A_f K_{f1-2}(T_{f_1} - T_{f_2}) - A_f K_{f2-3}(T_{f_2} - T_{f_3}) \tag{18}$$

$$\rho_{f_3} c_{f_3} A_f \delta_{f_3} \frac{\partial T_{f_3}}{\partial t} = A_f K_{f2-3}(T_{f_2} - T_{f_3}) - A_f K_{f3-S}(T_{f_3} - T_{soil}) \tag{19}$$

(3) Building heat demand

The heat consumption of the building includes the heat consumption generated by the heat convection between the indoor air and the inner surface and the heat consumption generated by the infiltration. Solar heat gain is the only indoor heat gain considered in this study, other kinds of heat gain such as personnel heat gain and equipment heat gain are ignored. In summary, the heat demand equation of the building is shown in Equation (20).

$$\begin{aligned} Q_d = h_{in}[A_{wall}(T_{in} - T_{wall,in}) + A_{roof}(T_{in} - T_{roof,in}) + A_{floor}(T_{in} - T_{floor,in})] + K_{glassi} A_{glass}(T_{in} - T_o) \\ +0.278 n_k V_{in} c_{in} \rho_{in}(T_{in} - T_o) - Q_{solar} \end{aligned} \tag{20}$$

where, $Q_d$ is the total heat demand of the building, W; $n_k$ is the air change rate, $n_k$ = 0.5 $h^{-1}$; $c_{in}$ is the heat capacity of the indoor air, $c_{in}$ = 1003 J/(kg·°C); $V_{in}$ is the volume of the building, $V_{in}$ = 33.6 $m^3$; $\rho_{in}$ is the indoor air density, $\rho_{in}$ = 1.29 kg/$m^3$, $Q_{solar}$ is the indoor solar heat gain, which is given in Equation (21).

$$Q_{solar} = A \times EF \times I_{sun} \times SHGC \tag{21}$$

where, $SHGC$ is the solar heat gain coefficient; $EF$ is the effective area factor.

(4) Dynamic equation for indoor temperature

The dynamic equation for indoor temperature can be calculated by Equation (22). The research objective of this study is dynamic on-demand heating operation, i.e., $Q_d - Q_s = 0$, and $\frac{dT_{in}}{dt} = 0$, hence the indoor temperature should remain stable.

$$V_{in} c_{in} \rho_{in} \frac{dT_{in}}{dt} = Q_d - Q_s \tag{22}$$

where, $Q_s$ is the heat supply to the building, W.

2.4.2. Reward Function

In the process of RL, the design of the reward function plays a very important role, because it will determine the convergence speed and final performance of the RL model. The agent always learns how to maximize the reward, so the reward function must be designed ensuring the consistency between maximizing rewards and objectives. It is crucial that the reward function is designed in a way that really shows the training objective.

Delayed reward is one of the most difficult problems in RL, because there is a lot of exploratory work that needs to be performed continuously and iteratively in the process of RL training. Take the go game as an example, the agent will only achieve a $-1/+1$ reward at the end of a game to judge whether to win or lose, which makes training extremely difficult. For the prediction of building heat demand, there is no conventional RL reward function design method [34]. Table 3 summarizes the common reward function design approaches.

**Table 3.** Summary of different methods for reward function design.

| Types of Reward Function | Characteristics |
| --- | --- |
| Discrete reward function [29] | Contains less information |
| Continuous reward function [35] | Contains more information, and suitable for continuous action space |
| Inverse reinforcement learning [36] | Require a large amount of labeled data |

In order to achieve the objective of on-demand heating operation, the most appropriate evaluation index of the heating performance through the building heat demand prediction is the indoor temperature. According to relevant standards in China, the indoor temperature in winter should be between 16 °C and 24 °C, however, this range is usually raised to 18 °C to 24 °C during actual operation considering the thermal comfort of residents. Therefore, in this study, the desired range for indoor temperature control is 18 °C to 24 °C, and the target temperature of indoor temperature control is the midpoint of this range, i.e., 21 °C. As a result, this study proposes the following three reward functions for the building heat demand prediction. The temperature is divided into 7 intervals, the interval 0.5 degrees above and below the target temperature is the target interval; the two sides of the target interval are suitable intervals; the outer side of the suitable intervals are qualified intervals; and the outer side of the qualified intervals are disqualified intervals. Both target interval and suitable interval are collectively referred to as acceptable interval. The basic principle of reward function design is that the closer it is to the target temperature, the higher the reward.

(1) Reward function 1 (R1): only the target interval has a slightly positive reward, as shown in Figure 5.
(2) Reward function 2 (R2): the reward of the target interval has the same absolute value of the reward of the disqualified intervals and the suitable interval has a slightly positive reward, as shown in Figure 6.
(3) Reward function 3 (R3): on the basis of R2, the reward of the target interval and the right side of the qualified interval and disqualified interval are halved because lower temperature is even more strongly disallowed than higher temperature, considering the thermal comfort of residents, as shown in Figure 7.

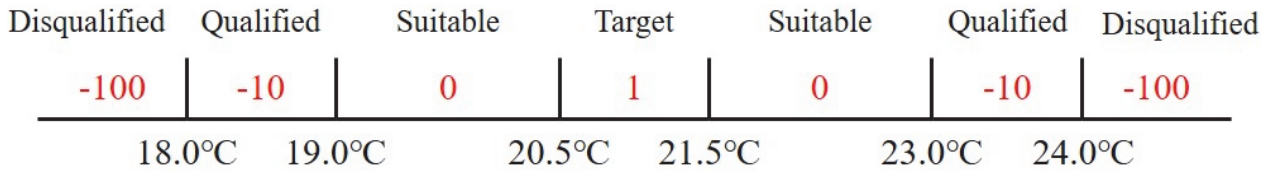



**Figure 5.** Design of reward function 1.

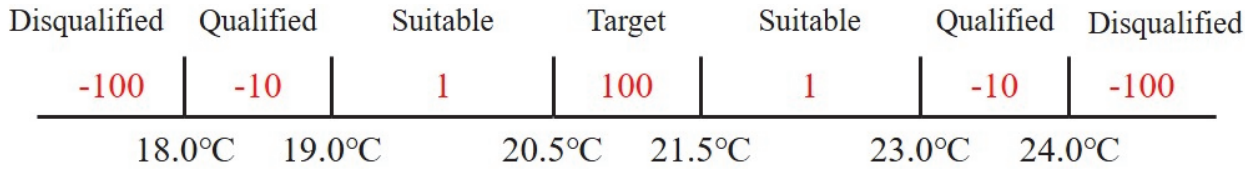



**Figure 6.** Design of reward function 2.

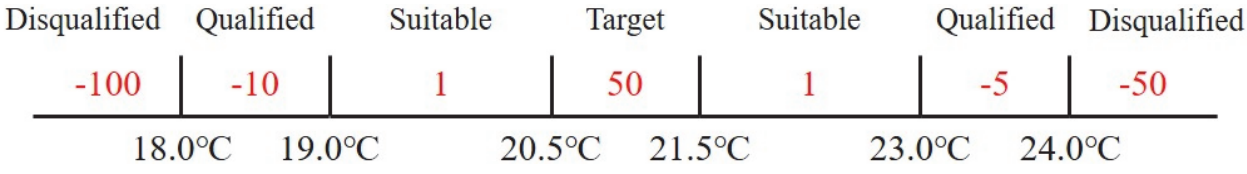



**Figure 7.** Design of reward function 3.

#### 2.4.3. Agent

Due to the fact that the output of building heat demand prediction is a high-dimensional continuous action space, this study builds the agent based on the TD3 algorithm. The goal of the agent is to find the best action under different state parameters through interaction with the environment, so that the highest cumulative reward can be obtained according to the reward function. Since the objective of this study is to achieve on-demand heating operation, the optimal action given by the agent is both the building heat demand and the heat supply of the system.

The neural networks of the agent proposed by this study is shown in Figure 8. The Actor outputs the strategy and is responsible for selecting actions; while the Critic is responsible for estimate the Q-value of each action. In the Actor, the input features at time t include indoor temperature at time t, heat supply at time t, outdoor temperature at time t + 1, and solar radiation intensity at time t + 1. The input features go through an LSTM layer and a fully connected (FC) layer for nonlinear transformation, and then through the Dropout layer to solve the problem of over-fitting, and finally reaches the output layer. The output is the predicted building heat demand at time t + 1, which is also the heat that needs to be supplied to the building at time t + 1. In the Critic, in addition to the input features required by the Actor, the input features also include the action selected by the Actor at time t + 1. The architecture of the Critic has two training branches. The left branch is designed for feature extraction of state observation, which is consist of a LSTM layer and a FC layer. While the right branch is designed for nonlinear transformation of corresponded actions, which only contains a LSTM layer. The outputs of the two branches are merged together through the merge layer. The merge layer uses the vector link strategy without weight parameters, because merge methods such as addition, subtraction or average summation will cause a certain degree of information loss. Finally, the dropout layer is added after the FC layer to solve the overfitting.

Combined with the environment, reward function and agent described previously, a complete building heat demand prediction model based on RL can be established. By observing the environment, the agent obtains the state parameters $S_t$, including indoor temperature, outdoor temperature and solar radiation intensity, and then predicts the heat supply action $a_t$. After the action is executed, the new state parameters $S_{t+1}$ and the reward $R_t$ can be given by the environment. Hence, the continuous interaction between the agent and the environment can be realized in this loop.

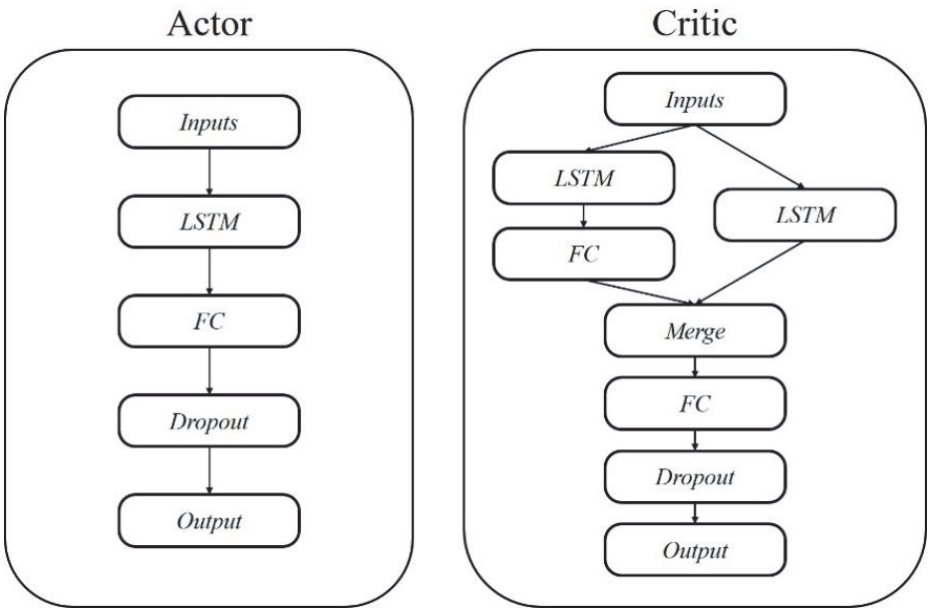



**Figure 8.** The architecture of the Actor and the Critic.

### *2.5. Experiment Schemes*

This study analyzed the training and prediction performance of the model based on the architecture proposed previously. Since the objective is to achieve on-demand heating operation on the premise of ensuring indoor comfort, this study will evaluate the indoor temperature control performance and the supply–demand error of proposed model based on building heat demand prediction to verify its feasibility and advancement.

#### 2.5.1. Model Settings

In the experiment, the agent needed to find the best heat supply corresponding to each step by interacting with the environment, so that the indoor temperature of the simulated indoor environment can be maintained at 21 °C. The total duration of the experiment was 7 days, and the interval of each step was 1 h, i.e., a complete episode took a total of 168 h. The training process will stop when the maximum number of the episode is reached or the average reward of the previous several episodes reach the preset value.

(1) Meteorological data

The meteorological data in the heating season of 2018–2019 were selected, including outdoor temperature and solar radiation intensity, and the data were in an interval of 1 h. The hourly outdoor temperature is shown in Figure 9, and the hourly solar radiation intensity is shown in Figure 10.

(2) Parameters settings for the neural networks

Considering that the delay of heat supply caused by thermal inertia of building envelopes is usually within 4 h, the sequence length of the LSTM was set to 4 in the simulation. The other parameters of the actor network and the critic network are shown in Tables 4 and 5, respectively.

(3) Parameters settings for the reinforcement learning

The main parameters for the RL are set as shown in Table 6.

All the algorithms were established and implemented by Python 3.6.7 and all the results in this study were obtained by a computer with a processor of 2.4 GHz Intel Core i7 and 8 GB RAM.

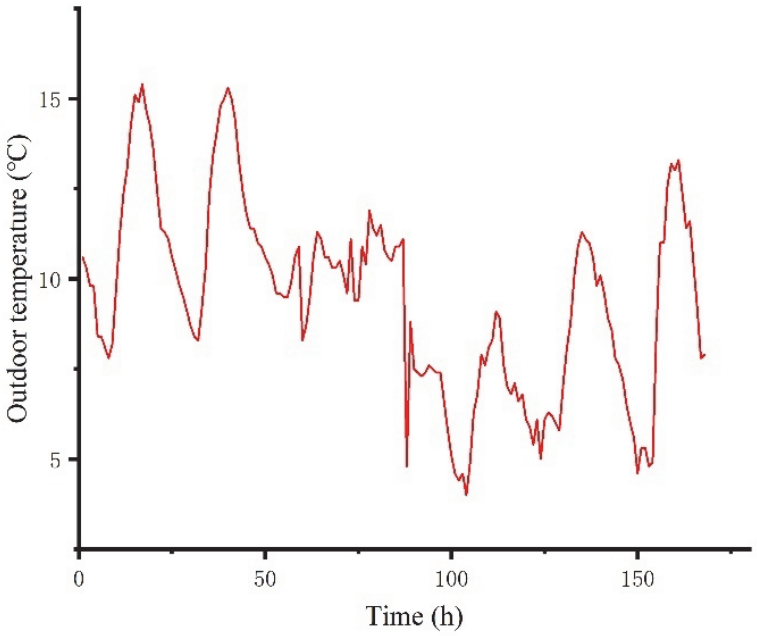



**Figure 9.** Hourly outdoor temperature.

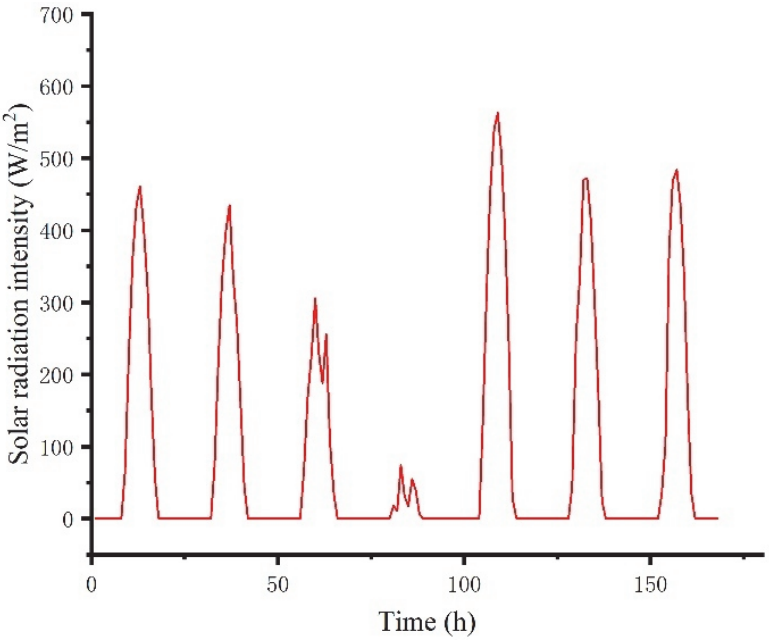



**Figure 10.** Hourly solar radiation intensity.

**Table 4.** Parameters settings for the Actor network.

| Parameters | Values |
|---|---|
| Neuron number of the LSTM | 256 |
| Activation function of the LSTM | ReLU |
| Neuron number of the FC | 128 |
| Activation function of the FC | ReLU |
| Optimizer | Adam |
| Learning rate | 0.001 |
| Dropout | 0.5 |

**Table 5.** Parameters settings for the Critic network.

| Parameters | Branch of State | Branch of Action |
|---|---|---|
| Neuron number of the LSTM | 256 | 128 |
| Activation function of the LSTM | ReLU | ReLU |
| Neuron number of the FC | 128 | - |
| Activation function of the FC | ReLU | - |
| Optimizer | Adam | Adam |
| Learning rate | 0.001 | 0.001 |
| Dropout | 0.5 | 0.5 |

**Table 6.** Parameters settings for reinforcement learning.

| Parameters | | Values |
|---|---|---|
| Sample time | | 1 |
| Discount factor | | 0.995 |
| Maximum episode | | 1000 |
| Average window length | | 10 |
| Exploration Method | Type of noise | Gaussian action noise |
| | Mean | 50 |
| | Standard deviation | 200 |
| | Decay of SD | 0.00001 |
| Batch size | | 64 |

#### 2.5.2. Evaluation Indices

As given in Equation (22), for on-demand heating operation, $Q_d - Q_s = 0$ and hence $\frac{dT_{in}}{dt} = 0$, which indicates that the indoor temperature should remain stable. Therefore, the difference between the actual indoor temperature and the target indoor temperature at each step is the optimal criteria to evaluate the heating performance based on building heat demand prediction. The smaller the difference, the better the heating performance.

This study used the Root Mean Squared Error (RMSE) and the Coefficient of Variation of the RMSE (CV-RMSE) as the evaluation indices to analyze the prediction performance of proposed model, as given in Equations (23) and (24). RMSE and CV-RMSE are widely used for model performance evaluation, 41% and 16% of the research studies select them as the model performance evaluation indices, respectively [37].

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(t_i - t_t)^2}{n}} \tag{23}$$

$$\mathrm{CV-RMSE} = \frac{\sqrt{\frac{\sum_{i=1}^{n}(t_i - t_t)^2}{n}}}{\frac{\sum_{i=1}^{n} t_i}{n}} \times 100\% \tag{24}$$

where, $t_i$ is the actual indoor temperature observed form the environment, °C; $t_t$ is the target temperature, $t_t$ = 21 °C; $n$ is the total steps in a single episode, $n$ = 168.

## 3. Results and Discussions

### *3.1. Model Training*

According to the model settings mentioned in Section 2.5.1, the three proposed reward functions were used to train the model. In addition, the training target related to the end of training, i.e., the average reward of the last 10 episodes, were set as primary, intermediate and advanced targets, corresponding to 1/3, 1/2 and 2/3 of the total reward, respectively. The result is shown in Figure 11.

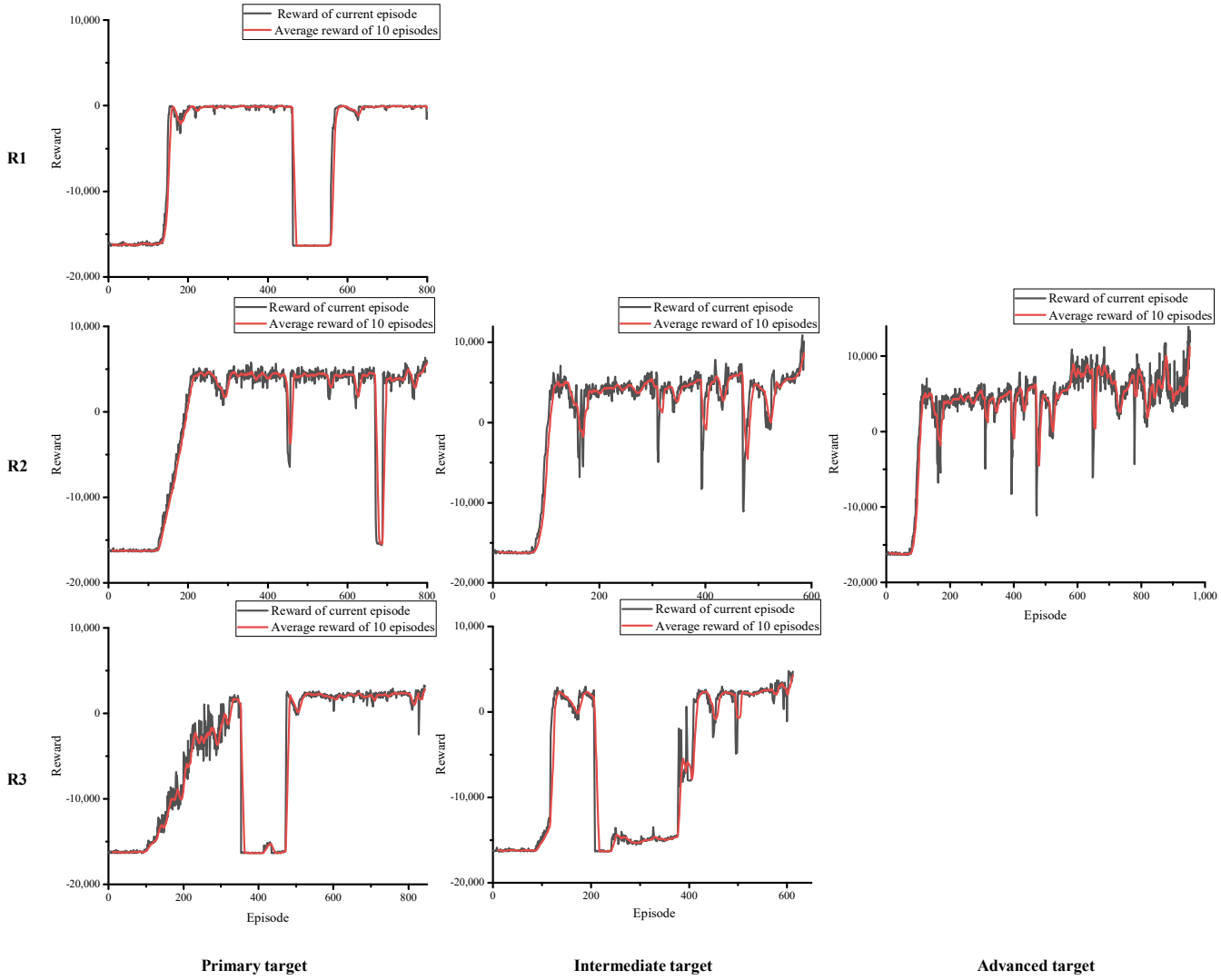



**Figure 11.** Training results of the three proposed reward functions.

The results show that when R1 was used, the highest average reward in the past 10 episodes is only −28.3, while the total reward of a single episode is 168, and the highest reward in a single episode is 53. After training, the primary target of 1/3 of the total reward was still not reached, so the model training failed. When R2 and R3 were used, both the primary and intermediate targets can be achieved. To reach the primary target, the target reward of R2 is 5600. After 800 episodes of training, the model achieved the primary target, and the average reward of 10 episodes was 5671. The primary target of R3 was 2800. After 845 episodes, the model achieved the primary target, and the average reward of 10 episodes was 2832. On the other hand, to reach the intermediate target, the target reward of R2 is 8400. After 586 episodes of training, the model achieved the intermediate target, and the average reward of 10 episodes was 8650. The intermediate target of R3 is 4200. After 612 episodes, the average reward of 10 episodes was 4264 which achieved the intermediate target. Aiming to achieve the advanced target, the model cannot be successfully trained directly using any of the three reward functions proposed in this study. However, if the pretrained model of R2 which achieved the intermediate target was used for retraining, the average reward of 10 episodes was 11,850 after 367 episodes which achieved the advanced target. In summary, results show that R2 is the most suitable for the experimental situation proposed in this study.

The experimental results show that the design of the reward function significantly influences the performance of the model after training. For the same architecture of the

agent, the model failed in training when R1 was used, while the model can achieve the advanced target when R2 was used through retraining. The reward distribution of R1 is very sparse due to the fact that there is only one positive reward interval in R1. As a result, the probability that the model can achieve a positive reward during training is low, and the model may not be able to find the correct optimization direction after many attempts, which seriously affects model training efficiency. In addition, the positive and negative rewards set in R1 are not uniform in magnitude, the positive reward is only 1, while the negative reward is −10 and −100. Consequently, the model is more likely to tend to avoid high negative rewards, rather than obtaining high positive rewards during training. The training direction is inconsistent with our training objective. This phenomenon is also verified in the training results, the rewards of most late episodes are small negative values rather than positive values, which indicates a design failure of R1.

Compared with R1, since the reward of target interval in R2 and R3 has the same order of magnitude as the disqualified intervals, the training performance is significantly improved. In addition, the reward of suitable intervals in R2 and R3 is set to 1, which aims to guide the optimization direction of model training. The experimental results also verify the effectiveness of this method. From the perspective of indoor thermal comfort, residents are more reluctant to accept indoor temperatures below 18 °C compared to above 24 °C, because high indoor temperature can be decreased by opening the windows. Therefore, R3 halves the reward for target interval and high temperature intervals, aiming to guide the model to avoid low indoor temperature during training. However, the experimental results show that using R3 does not have better performance, it fails to reach the advanced target even retraining method is used.

With the upgrade of model training target, the difficulty of model training will also increase significantly. Due to all the default parameters of neural network model are randomly initialized at the beginning, the model proposed in this study is not always successful during training. It is found that the training will always fail when R1 is used. However, when the other reward functions are used, the success probabilities of the primary target and the intermediate target are about 40% and 15%, respectively. When R2 is used by the retraining method, the success probability to achieve the advanced target is about 20%. In addition, due to random initialization, even the same agent architecture and reward function are used, the parameters of the final model obtained from each training attempt are different. Therefore, it is not meaningful to compare the number of episodes used for each training attempt. For example, during two different training attempts with R2, it took 800 episodes to reach the primary target, but only took 586 episodes to reach the intermediate target.

### *3.2. The Performance of Indoor Temperature Control*

Figure 12 shows the variation of indoor temperature when the environment was controlled by different prediction models.

When R2 was used, the model which achieved primary target could control the average indoor temperature within 168 h of a complete single episode to 21.6 °C, the RMSE is 1.3 °C and the CV-RMSE is 6.26%. The indoor temperature was controlled within the target interval for 60 h, accounting for 35.7%, within the acceptable interval for 156 h, accounting for 92.9%, and within the qualified interval for 12 h, accounting for 7.1%. The model achieved intermediate target can control the average indoor temperature of a complete single episode to 20.8 °C, the RMSE is 0.5 °C and the CV-RMSE is 2.67%. The indoor temperature was controlled within the target interval for 101 h, accounting for 60.1%, within the acceptable interval for 162 h, accounting for 96.4%, and within the qualified interval for 6 h, accounting for 3.6%. The model which achieved the advanced target could control the average indoor temperature of a complete single episode to 21.2 °C, the RMSE is 0.3 °C and the CV-RMSE is 1.23%. The indoor temperature was controlled within the target interval for 118 h, accounting for 70.2%, within the acceptable interval for 168 h, accounting for 100%.

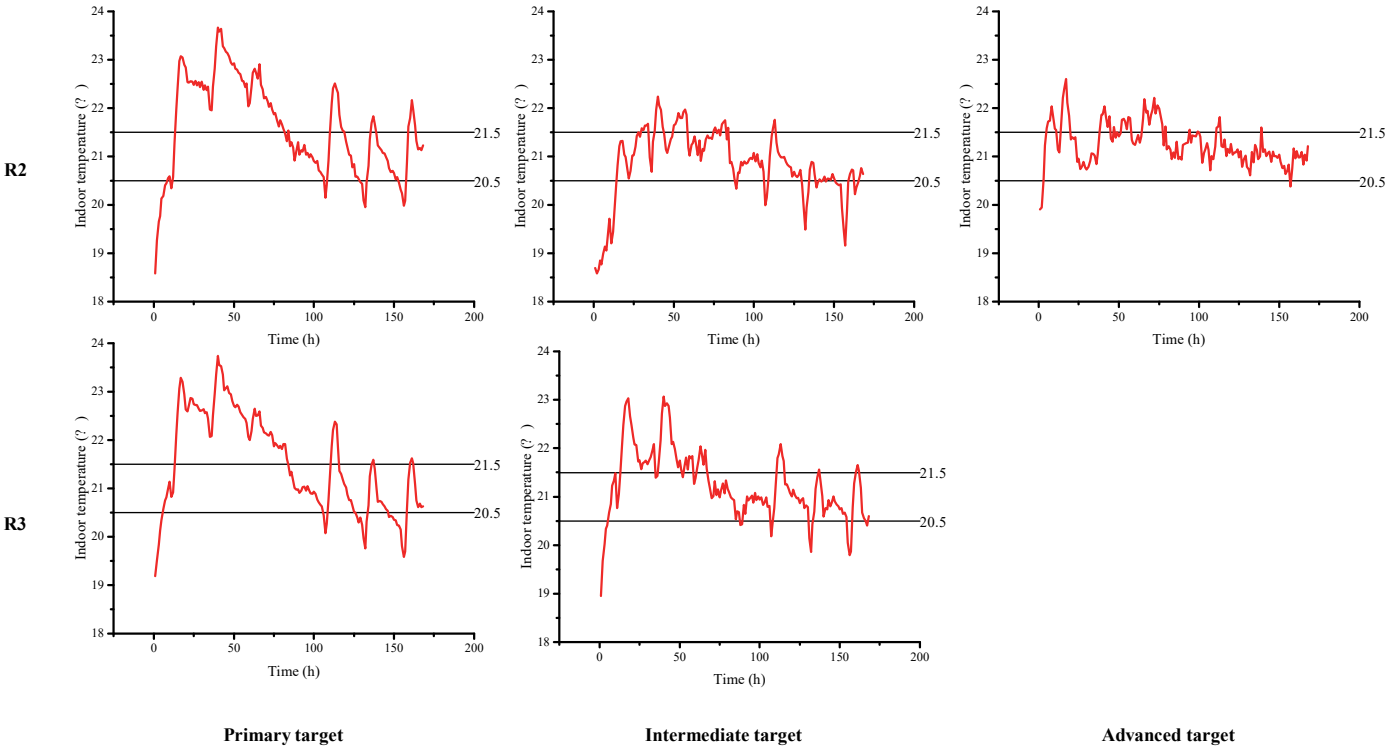



**Figure 12.** Results of indoor temperature controlled by different prediction models.

When R3 was used, the model which achieved the primary target could control the average indoor temperature within a complete single episode to 21.5 °C, the RMSE is 1.3 °C and the CV-RMSE is 6.19%. The indoor temperature was controlled within the target interval for 62 h, accounting for 36.9%, within the acceptable interval for 157 h, accounting for 93.5%, and within the qualified interval for 11 h, accounting for 6.5%. The model which achieved the intermediate target could control the average indoor temperature of a complete single episode to 21.2 °C, the RMSE is 0.5 °C and the CV-RMSE is 2.61%. The indoor temperature was controlled within the target interval for 93 h, accounting for 55.4%, within the acceptable interval for 165 h, accounting for 98.2%, and within the qualified interval for 3 h, accounting for 1.8%.

The experimental results show that under the same training target, models trained with different reward functions have similar performance on indoor temperature control. When the primary target is achieved, the time for the indoor temperature in the target interval is more than 35%; when the intermediate target is achieved, this time increases to more than 55%; and it can further increase to more than 70% when the advanced target is achieved. The performance of indoor temperature control can be significantly improved with the upgrade of the training target, but it should be noticed that the upgrade of the training target will also lead to a significant increase in the difficulty of the model training. It can be seen in Figure 12 that even with the primary target, the model can maintain the indoor temperature in the acceptable interval for more than 90% of the time. All of the poor indoor temperature control occurs within the first 10 hours of the simulation. The reason is that the initial temperature of the internal node of each building envelope layer was set to 18 °C. Therefore, the heating scenario of the early stage is different from common operation, which requires additional preheating of the building envelopes. It is difficult to learn the preheating scenario in the training process, as a result, it is found in the experiment that the model cannot achieve advanced targets through conventional training methods. However, when the model achieves the intermediate target, the indoor temperature can be appropriately controlled except for the preheating scenario in the early stage. Therefore, on this basis, the retraining method can be used to fine-tune the original model and help the

model learn the characteristics of the preheating scenario. The experimental results also verify the effectiveness of retraining method.

### *3.3. The Performance of On-Demand Heating Operation*

Figure 13 shows the variation of heat supply when the environment was controlled by different prediction models.

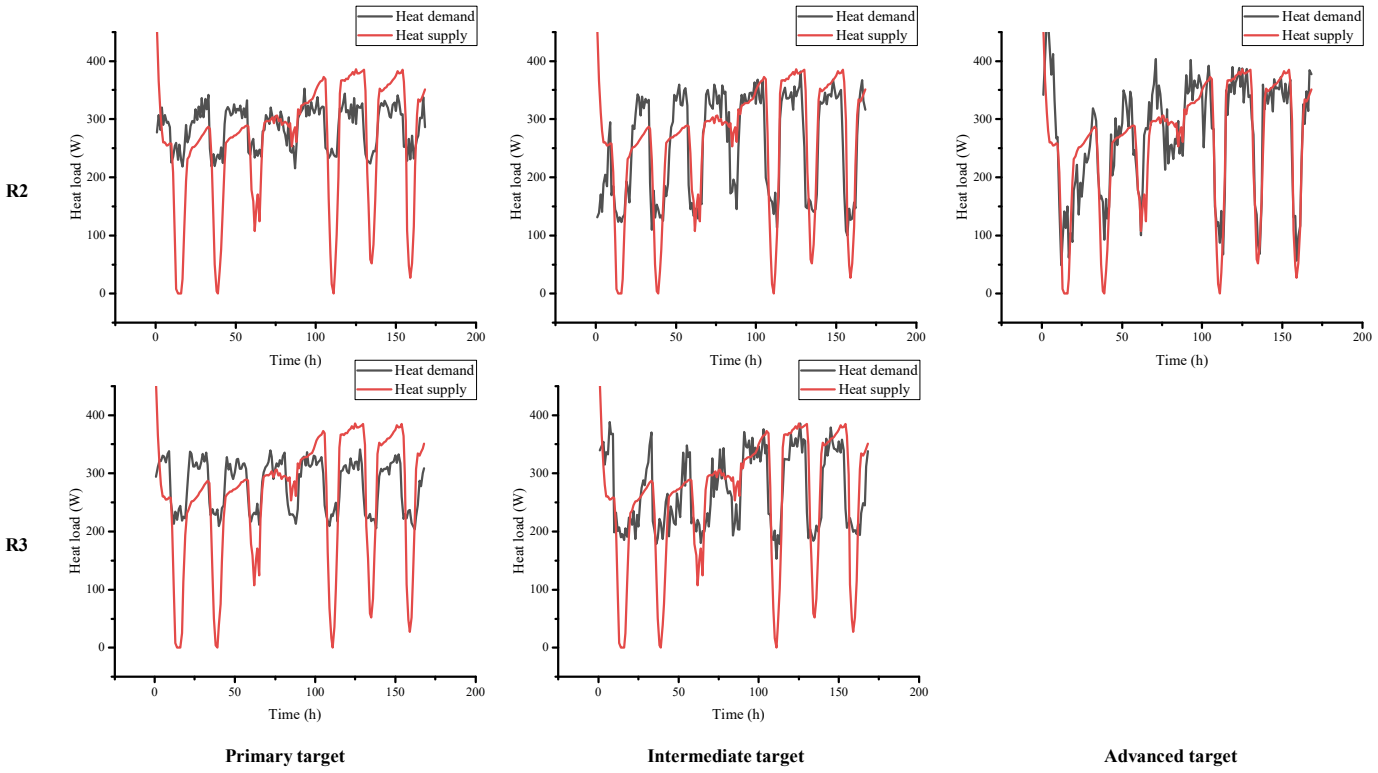



**Figure 13.** Results of heat supply controlled by different prediction models.

In the experiment, the total heat demand in a complete episode of 168 h is 43,665.4 kWh. When R2 was used, the total heat supply of the model that achieved the primary target, the intermediate target and the advanced target were 48,060.8 kWh, 42,804.4 kWh and 45,656.6 kWh, respectively. The supply–demand error is 10.07%, −1.97% and 4.56%, respectively. On the other hand, when R3 was used, the total heat supply of the model that achieved the primary target and the intermediate target is 47,148.8 kWh and 45,460.7 kWh, respectively. The supply–demand error is 7.98% and 4.11%, respectively.

Due to the difference in the average indoor temperature controlled by different models, the total heat supply in a single episode is slightly different. In general, the trained model can control the supply–demand error within 10%. For the model with the best indoor temperature control performance, the supply–demand error is only 4.56%. In addition, it can be found from Figure 13 that the indoor heat demand changes periodically. With the upgrade of the training target, the fitting of the heat supply and the heat demand curve is significantly improved. Based on the primary target, although the heat supply is consistent with the heat demand in the changing trend, the peaks and valleys cannot be well fitted. Under this condition, the indoor temperature can be maintained in the acceptable interval based on the thermal storage of building envelopes. However, the indoor temperature will still be very high during the low heat demand at noon every day. When the intermediate target is achieved in training, the previous problem can be significantly alleviated as the prediction accuracy is sharply improved. When training with the advanced target, with the introduction of the retraining method, the fitting of heat demand and heat supply is

further improved, and the problem of additional heat demand caused by the preheating of building envelopes in the early stage can be solved.

## 4. Conclusions

The accurate prediction of heat demand plays the critical role in refined management of heating, which is the basis for on-demand heating operation. However, the model established by the supervised learning method requires the operating conditions of the training data realized on-demand heating in advance, in order to achieve its on-demand heating operation. This study proposed a prediction model framework based on reinforcement learning according to indoor temperature to precisely predict the heat demand and achieve on-demand heating operation. This study established the agent based on the TD3 algorithm to dynamically maintain the indoor temperature at around 21 °C. A simulation environment was constructed based on building physical parameters and mathematical model to cooperate with the agent for training. In addition, three reward functions for the building heat demand prediction model were elaborate based on the indoor temperature considering practical heating performance. The main conclusions are as follows:

(1) The experimental results showed that the design of the reward function significantly influenced the performance of the model after training. The agent failed to achieve the primary target that 1/3 of the total reward based on R1. However, the agent achieved the advanced target and intermediate target, respectively, when R2 and R3 were used.
(2) The model proposed in this study can dynamically control the indoor temperature in the acceptable interval through building heat demand prediction. The experimental results showed that even if the model only reached the primary target through training, it could maintain the indoor temperature in the acceptable interval (19–23 °C) for more than 90% of the time during the experiment. Moreover, the proportion of time that the indoor temperature can be controlled within the target interval (20.5–21.5 °C) was over 35%, 55% and 70%, when the model reached the primary, intermediate or advanced targets, respectively. This study verified the prediction accuracy of building heat demand from the perspective of indoor temperature control performance, and validated the effectiveness and advancement the model proposed in this study.
(3) In addition to maintaining indoor temperature, the model proposed in this study also achieved on-demand heating operation. Although the final heat supply was directly influenced by the indoor temperature control, the experimental results show that the supply–demand error can be controlled below 10% by training a model that achieved the primary target. The model achieving the advanced target, which had the best indoor temperature control performance, had a supply–demand error of only 4.56%, providing basis and support for the realization of on-demand heating operation.

**Author Contributions:** Conceptualization, C.W.; Data curation, C.W.; Formal analysis, C.W. and J.Y.; Investigation, L.Z. and K.H.; Methodology, C.W. and J.Y.; Project administration, Z.Z.; Software, C.W.; Supervision, Z.Z.; Validation, L.Z., J.Y. and K.H.; Visualization, L.Z. and K.H.; Writing—original draft, C.W. All authors have read and agreed to the published version of the manuscript.

**Funding:** We acknowledge financial support from Tianjin Science and Technology Commission (No. 21JCZDJC00660).

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. The State Council Information Office of the People's Republic of China. Responding to Climate Change: China's Policies and Actions. 2021. Available online: http://www.scio.gov.cn/ztk/dtzt/44689/47315/index.htm (accessed on 8 September 2022). (In Chinese)
2. Yuan, J.; Huang, K.; Han, Z.; Wang, C.; Lu, S.; Zhou, Z. Evaluation of the operation data for improving the prediction accuracy of heating parameters in heating substation. *Energy* **2022**, *238*, 121632. [CrossRef]
3. Tsinghua University Building Energy Research Center. *Annual Report on China Building Energy Efficiency*; China Construction Industry Publishing House: Guangzhou, China, 2021.

4. Zhang, Y.; Wang, J.; Chen, H.; Zhang, J.; Meng, Q. Thermal comfort in naturally ventilated buildings in hot-humid area of China. *Build. Environ.* **2010**, *45*, 2562–2570. [CrossRef]
5. Zheng, X.; Wei, C.; Qin, P.; Guo, J.; Yu, Y.; Song, F.; Chen, Z. Characteristics of residential energy consumption in China: Findings from a household survey. *Energy Policy* **2014**, *75*, 126–135. [CrossRef]
6. Dotzauer, E. Simple model for prediction of loads in district-heating systems. *Appl. Energy* **2002**, *73*, 277–284. [CrossRef]
7. Kuster, C.; Rezgui, Y.; Mourshed, M. Electrical load forecasting models: A critical systematic review. *Sustain. Cities Soc.* **2017**, *35*, 257–270. [CrossRef]
8. Yuan, J.; Wang, C.; Zhou, Z. Study on refined control and prediction model of district heating station based on support vector machine. *Energy* **2019**, *189*, 116193. [CrossRef]
9. Wang, R.; Lu, S.; Feng, W. A novel improved model for building energy consumption prediction based on model integration. *Appl. Energy* **2020**, *262*, 114561. [CrossRef]
10. Wang, R.; Lu, S.; Li, Q. Multi-criteria comprehensive study on predictive algorithm of hourly heating energy consumption for residential buildings. *Sustain. Cities Soc.* **2019**, *49*, 101623. [CrossRef]
11. Chae, Y.T.; Horesh, R.; Hwang, Y.; Lee, Y.M. Artificial neural network model for forecasting sub-hourly electricity usage in commercial buildings. *Energy Build.* **2016**, *111*, 184–194. [CrossRef]
12. Fan, C.; Sun, Y.; Zhao, Y.; Song, M.; Wang, J. Deep learning-based feature engineering methods for improved building energy prediction. *Appl. Energy* **2019**, *240*, 35–45. [CrossRef]
13. Xue, G.; Pan, Y.; Lin, T.; Song, J.; Qi, C.; Wang, Z. District heating load prediction algorithm based on feature fusion LSTM model. *Energies* **2019**, *12*, 2122. [CrossRef]
14. Song, J.; Zhang, L.; Xue, G.; Ma, Y.; Gao, S.; Jiang, Q. Predicting hourly heating load in a district heating system based on a hybrid CNN-LSTM model. *Energy Build.* **2021**, *243*, 110998. [CrossRef]
15. Kim, T.; Cho, S. Predicting residential energy consumption using CNN-LSTM neural networks. *Energy* **2019**, *182*, 72–81. [CrossRef]
16. Tai, L.; Paolo, G.; Liu, M. Virtual-to-real deep reinforcement learning: Continuous control of mobile robots for mapless navigation. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Vancouver, BC, Canada, 24–28 September 2017; pp. 31–36.
17. Behzadan, V.; Munir, A. Vulnerability of deep reinforcement learning to policy induction attacks. In *International Conference on Machine Learning and Data Mining in Pattern Recognition*; Springer: Cham, Switzerland, 2017; pp. 262–275.
18. Zhao, D.; Chen, Y.; Lv, L. Deep reinforcement learning with visual attention for vehicle classification. *IEEE Trans. Cogn. Dev. Syst.* **2016**, *9*, 356–367. [CrossRef]
19. Liu, G.; Schulte, O.; Zhu, W.; Li, Q. Toward interpretable deep reinforcement learning with linear model u-trees. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*; Springer: Cham, Switzerland, 2018; pp. 414–429.
20. Yu, C.; Yang, T.; Zhu, W.; Li, G. Learning shaping strategies in human-in-the-loop interactive reinforcement learning. *arXiv* **2018**, arXiv:1811.04272.
21. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G.; et al. Human-level control through deep reinforcement learning. *Nature* **2015**, *518*, 529–533. [CrossRef]
22. Tso, S.; Liu, K. Hidden markov model for intelligent extraction of robot trajectory command from demonstrated trajectories. In Proceedings of the IEEE International Conference on Industrial Technology, Shanghai, China, 2–6 December 1996; pp. 294–298.
23. Kruger, V.; Herzog, D.L.; Baby, S.; Ude, A.; Kragic, D. Learning actions from observations. *IEEE Robot. Autom. Mag.* **2010**, *17*, 30–43. [CrossRef]
24. Liu, T.; Tan, Z.; Xu, C.; Chen, H.; Li, Z. Study on deep reinforcement learning techniques for building energy consumption forecasting. *Energy Build.* **2020**, *208*, 109675. [CrossRef]
25. Mocanu, E.; Mocanu, D.C.; Nguyen, P.H.; Liotta, A.; Webber, M.E.; Gibescu, M.; Slootweg, J.G. On-line building energy optimization using deep reinforcement learning. *IEEE Trans. Smart Grid* **2018**, *10*, 3698–3708. [CrossRef]
26. Mason, K.; Grijalva, S. A review of reinforcement learning for autonomous building energy management. *Comput. Electr. Eng.* **2019**, *78*, 300–312. [CrossRef]
27. Baghaee, S.; Ulusoy, I. User Comfort and Energy Efficiency in HVAC Systems by Q-learning. In Proceedings of the Signal Processing and Communications Applications Conference (SIU), Izmir, Turkey, 2–5 May 2018; pp. 1–4.
28. Faddel, S.; Tian, G.; Zhou, Q.; Aburub, H. Data Driven Q-Learning for Commercial HVAC Control. In Proceedings of the SoutheastCon, Raleigh, NC, USA, 28–29 March 2020.
29. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An introduction*; MIT Press: Cambridge, MA, USA, 2018.
30. Lillicrap, T.P.; Hunt, J.J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; Wierstra, D. Continuous control with deep reinforcement learning. *arXiv* **2015**, arXiv:1509.02971.
31. Tang, H.; Houthooft, R.; Foote, D.; Stooke, A.; Xi Chen, O.; Duan, Y.; Schulman, J.; De Turck, F.; Abbeel, P. Exploration: A study of count-based exploration for deep reinforcement learning. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017; Volume 30, pp. 1–18.
32. Plappert, M.; Houthooft, R.; Dhariwal, P.; Sidor, S.; Chen, R.Y.; Chen, X.; Asfour, T.; Abbeel, P.; Andrychowicz, M. Parameter space noise for exploration. *arXiv* **2017**, arXiv:1706.01905.
33. Fujimoto, S.; Hoof, H.; Meger, D. Addressing function approximation error in actorcritic methods. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 1587–1596.

34. Ng, A.Y.; Harada, D.; Russell, S. Policy invariance under reward transformations: Theory and application to reward shaping. *Icml* **1999**, *99*, 278–287.
35. Devidze, R.; Radanovic, G.; Kamalaruban, P.; Singla, A. Explicable reward design for reinforcement learning agents. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 20118–20131.
36. Balakrishnan, S.; Nguyen, Q.P.; Low BK, H.; Soh, H. Efficient Exploration of Reward Functions in Inverse Reinforcement Learning via Bayesian Optimization. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 4187–4198.
37. Amasyali, K.; El-Gohary, N.M. A review of data-driven building energy consumption prediction studies. *Renew. Sustain. Energy Rev.* **2018**, *81*, 1192–1205. [CrossRef]