

Article

System Frequency Control Method Driven by Deep Reinforcement Learning and Customer Satisfaction for Thermostatically Controlled Load

Rusi Chen ¹, Haiguang Liu ¹, Chengquan Liu ^{2,*}, Guangzheng Yu ², Xuan Yang ³ and Yue Zhou ³¹ State Grid Hubei Electric Power Research Institute, Wuhan 430077, China² Department of Electrical Engineering, Shanghai University of Electric Power, Shanghai 201306, China³ State Grid Hubei Electric Power Co., Ltd., Wuhan 430072, China

* Correspondence: y20103028@mail.shiep.edu.cn

Abstract: The intermittence and fluctuation of renewable energy aggravate the power fluctuation of the power grid and pose a severe challenge to the frequency stability of the power system. Thermostatically controlled loads can participate in the frequency regulation of the power grid due to their flexibility. Aiming to solve the problem of the traditional control methods, which have limited adjustment ability, and to have a positive influence on customers, a deep reinforcement learning control strategy based on the framework of soft actor–critic is proposed, considering customer satisfaction. Firstly, the energy storage index and the discomfort index of different users are defined. Secondly, the fuzzy comprehensive evaluation method is applied to evaluate customer satisfaction. Then, the multi-agent models of thermostatically controlled loads are established based on the soft actor–critic algorithm. The models are trained by using the local information of thermostatically controlled loads, and the comprehensive evaluation index fed back by users and the frequency deviation. After training, each agent can realize the cooperative response of thermostatically controlled loads to the system frequency only by relying on the local information. The simulation results show that the proposed strategy can not only reduce the frequency fluctuation, but also improve customer satisfaction.

Keywords: thermostatically controlled load; frequency regulation; customer satisfaction; soft actor–critic; energy storage index; discomfort index



Citation: Chen, R.; Liu, H.; Liu, C.; Yu, G.; Yang, X.; Zhou, Y. System Frequency Control Method Driven by Deep Reinforcement Learning and Customer Satisfaction for Thermostatically Controlled Load. *Energies* **2022**, *15*, 7866. <https://doi.org/10.3390/en15217866>

Academic Editors: Luis Hernández-Callejo, Sergio Nesmachnow and Sara Gallardo Saavedra

Received: 18 September 2022

Accepted: 18 October 2022

Published: 24 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

With the increasing proportion of renewable energy in the power grid, the characteristics of intermittence and fluctuation will bring considerable challenges to the active power balance and frequency stability of the power grid [1]. The traditional power system maintains the balance of the system by adjusting the output of the generating side units. The regulation method is relatively simple and will generate additional economic and environmental costs [2]. In addition, with the increase in power load and the extensive access to renewable energy, the regulation capacity of the power generation side gradually decreases [3]. The power system with renewable energy as the main body can utilize advanced information technology to integrate and dispatch demand-side resources to provide a variety of auxiliary services [4,5]. Therefore, reasonable control of demand-side resources can supplement the traditional system frequency regulation, and thus enhance the stability of the power system [6].

In the demand-side resources, the thermostatically controlled load (TCL) is a kind of electric equipment controlled by a thermostat, which can realize electric heating conversion and adjustable temperature, including in heat pumps, electric storage water heaters (ESWHs), refrigerators, and heating, ventilation and air conditioning (HVAC) systems [7]. TCL can be used to provide frequency regulation services, and is mainly based on the

following three points. Firstly, it is widely distributed in residential, commercial and industrial buildings, with adjustable potential. Secondly, it has sufficient thermal storage capacity and can be regarded as distributed energy storage equipment. Thirdly, the control method is flexible and can respond to the power demand of the system in time [8]. Therefore, in order to fully excavate the frequency regulation potential of flexible resources on the demand side and maintain the grid frequency within a certain offset range, it is necessary to conduct in-depth research on the control strategy of large-scale TCLs on the demand side.

In the current research, there are mainly three control methods for TCLs to participate in ancillary services: centralized control, decentralized control and hybrid control [9–11]. In centralized control, the control center sends control signals to all controlled loads, but it needs to build a large number of communication channels, leading to high control costs. Hu et al. [9] established a hierarchical centralized load tracking control framework that coordinates demand-side heterogeneous TCL aggregators and uses a state-space model for modeling. The decentralized control decentralizes the judgment mechanism of load control to the local control terminal, and pre-sets the procedures or thresholds at the local control terminal. When the demand-side device detects important parameter changes, the load acts according to the pre-set strategy. Because the judgment of decentralized control is performed at the local port, the demand for communication is low and the response speed is fast. However, the control effect is largely influenced by the user behavior and the error of the detection device. Delavari and Kamwa [10] applied a multi-objective optimization approach for optimizing each load setting to reduce the amount of load response required and trigger the load based on the frequency response index of decentralized control. The hybrid control combines the features of centralized and decentralized control, and establishes a control framework of “centralized parameter setting–decentralized decision making”, and coordinates large-scale users and grid control centers through load aggregators (LAs). Song et al. [11] built a two-stage control model based on hybrid control to participate in energy market trading. Based on hybrid control, Wang et al. [12] used TCLs to mitigate PV and load variations in microgrid communities. The above methods require a communication network between the control center and all aggregates, increasing the cost and difficulty of demand-side load control.

In the research on the participation of TCLs in auxiliary service, Ref. [13] built a dynamic model and verified the performance of a variable-frequency heat pump in providing frequency modulation services by using direct load control. This paper mainly studied the dynamic response performance of a single air-conditioning system, but focused less on the coordinated control of large-scale air-conditioning loads. Ref. [14] established a virtual energy storage model of variable-frequency air conditioning, shielded part of the model information through a hierarchical control framework and simplified the downlink control by using a unified broadcast signal. However, in this paper, the adjustable capacity of the air conditioning cluster will be sacrificed in order to simplify the downlink control. There are two main control modes of TCL, namely direct switch and temperature setting [15]. Ref. [16] realized frequency adjustment based on load direct switch. The advantage of this method is that the tracking accuracy of the system is high and the influence on the user's comfort is low within the range of load regulation ability. The disadvantage is that when the indoor temperature of the load is concentrated near the temperature boundary, the equipment will be frequently switched on and off, which will not only fail to complete the adjustment task, but also reduce the service life of the equipment [17]. Temperature setting can avoid the above disadvantages, but its limitation is that the tracking effect of power depends on the designed controller including a minimum variance controller [18], a sliding mode controller [19] and an internal model controller [20]. In addition, its limitations are also shown in the large range of temperature changes, which will have an impact on the comfort of users [21]. Pallonetto et al. [22] established a residential building energy management system (EMS) based on a combination of optimization techniques and machine learning models. This EMS reduces energy consumption while maintaining thermal comfort. Therefore, it is important to consider the influence of consumer satisfaction in

the process of load response control to dispatch the motivation of user participation in demand response.

The above-mentioned documents all adopt a single control method to schedule and control TCL, but they often fail to meet the application requirements. This is not only because of their inherent defects, but also because of the different requirements of users. Direct switching is suitable for loads with high temperature requirements. At this time, the user expects the start and stop frequency of the equipment to be kept at a low level to extend the service life of the equipment. The load with a low temperature requirement is suitable for temperature setting. At this time, the user expects the temperature change to be maintained within a certain range to improve the comfort. Therefore, it is of practical significance to combine the two control modes. Ref. [23] proposed a hybrid control strategy based on a parallel structure. This strategy can improve the tracking accuracy of the system and reduce the number of switches in the equipment. However, the temperature changes widely, which will reduce the user's comfort. In recent years, the deep reinforcement learning algorithm has provided a new solution to the frequency control problem of power system.

With its strong search and learning ability, the deep reinforcement learning algorithm has the potential of online optimization of decision making in the face of complex nonlinear frequency control problems. The Q-learning algorithm of deep reinforcement learning is used to realize the cooperative control of distributed generation units, thus eliminating the frequency deviation of the system in [24,25]. Ref. [26] combined electric water heater buffer models with domain randomization to reduce the initialization time of Q-learning in demand response control. Ruelens et al. [27] applied batch reinforcement learning to coordinate the power consumption of users with thermostatically controlled loads. However, the Q-learning algorithm can only discretely select control actions from low-dimensional action domains, so it cannot deal with problems with continuous variables [28]. Ref. [29] proposed a deep reinforcement learning algorithm that acts on the continuous action domain, thus realizing the adaptive control of load frequency. An energy management scheme for AC control based on the deep deterministic policy gradient (DDPG) algorithm is proposed in [30]. However, this algorithm is only suitable for the optimal control of a single generator set, and is not suitable for the control of large-scale thermostatically controlled load. In [31], a distributed soft actor-critic (DSAC)-based data-driven frequency control method is proposed; the DSAC model estimates the distribution of value function over returns instead of only estimating the mean. The DSAC method based on entropy regularization has a faster learning speed compared to the traditional expectation-based reinforcement learning methods.

In view of the above problems, we take the thermostatically controlled load as the frequency control object, and based on the deep reinforcement learning algorithm, propose a frequency stability control method with the participation of thermostatically controlled load, considering customer satisfaction. Firstly, considering the operation characteristics of thermostatically controlled load with different control types, the energy storage index and the discomfort index are established, and the fuzzy comprehensive discrimination method is used to evaluate the customer satisfaction. Then, in order to realize the frequency cooperative control of large-scale thermostatically controlled load, a multi-agent control model is established based on the soft actor-critic algorithm to realize the continuous action control of thermostatically controlled load. Through the multi-agent reinforcement learning model considering customer satisfaction, the frequency response control of each thermostatically controlled load cluster can be coordinated online. The main contributions of this paper are as follows:

1. The influence of consumer satisfaction on the frequency response of TCLs is considered, and indicators reflecting satisfaction for TCLs with different control modes are established.
2. For the control problem of frequency response control for large-scale TCLs, this paper proposes a deep reinforcement learning control method based on the SAC algorithm.

This method can rely only on local load and frequency data to achieve real-time cooperative control of large-scale TCLs, which reduces the communication pressure in the scheduling process.

The remainder of this paper is organized as follows. In Section 2, the TCL dynamic model and control methods are formulated. Section 3 is concerned with the comprehensive control index of TCLs considering customer satisfaction. In Section 4, the frequency response control of TCLs based on SAC deep reinforcement learning is modeled. Case studies are provided in Section 5. Finally, conclusions are summarized in Section 6.

2. TCL Dynamic Model and Control Methods

2.1. TCL Dynamic Model

The first-order ordinary differential equation model considering indoor environment, outdoor environment and building characteristics has high accuracy and simple calculation, and is widely used in practice [32,33]. State variables T_i and virtual variables s_i are introduced into the model. The operating characteristics of the i -th TCL in the cooling mode can be expressed as:

$$\frac{dT_i(k)}{dk} = \frac{1}{C_i R_i} (T_\infty(k) - T_i(k) - s_i(k) R_i P_i) \quad (1)$$

Among them, the change rule of $s_i(k)$ is as follows:

$$s_i(k + \Delta k) = \begin{cases} 0 & s_i(k) = 1 \text{ and } T_i(k) \leq T_i^{\min} \\ 1 & s_i(k) = 0 \text{ and } T_i(k) \geq T_i^{\max} \\ s_i(k) & \text{others} \end{cases} \quad (2)$$

$$\begin{cases} T_i^{\min} = T_i^{\text{set}} - \frac{\delta}{2} \\ T_i^{\max} = T_i^{\text{set}} + \frac{\delta}{2} \end{cases} \quad (3)$$

where $T_\infty(k)$ and $T_i(k)$ are outdoor temperature and indoor temperature, respectively; C_i , P_i and R_i are the equivalent heat capacity, equivalent heat resistance and energy transfer rate of the i -th TCL, respectively; $s_i(k)$ indicates the load switch state, on state $s_i(k) = 1$ and off state $s_i(k) = 0$. $T_i^{\max}$ and $T_i^{\min}$ are the upper and lower limits of the temperature during load operation, respectively. T_i^{set} is the temperature setting value, δ is the temperature dead zone interval and is a constant. k and Δk are operation time and control period, respectively. By solving Equation (1):

$$T_i(k) = T_\infty(k) - s_i(k) R_i P_i - (T_\infty(k) - s_i(k) R_i P_i - T_i(0)) e^{-\frac{k}{C_i R_i}} \quad (4)$$

where $T_i(0)$ represents the initial indoor temperature. For a load cluster composed of N TCLs, the aggregate power consumption is the sum of the rated power of all loads.

$$P_{\text{total}}(k) = \sum_{i=1}^N P_i^n s_i(k) \quad (5)$$

$$P_i^n = \frac{P_i}{\eta_i} \quad (6)$$

where P_i^n is the rated power of the i -th TCL and η_i is the energy conversion efficiency coefficient of the i -th TCL.

Figure 1 shows the frequency response model of the power system with TCL load participating in frequency modulation, where T_{Ga} and T_{Gb} are the time constants of the governor and the turbine, respectively. The governor and the turbine are the instantaneous characteristic compensation links, which are expressed by the lead lag transfer function between the time constants T_1 and T_2 . T_R is the TCL response delay time constant, T_c is the

communication delay time constant and R_{eq} is the unit adjustment rate. ΔP_G , ΔP_L , H and D represent the total output power, the disturbance power, inertia time constant and load damping coefficient of the system, respectively. Δf is the frequency offset.

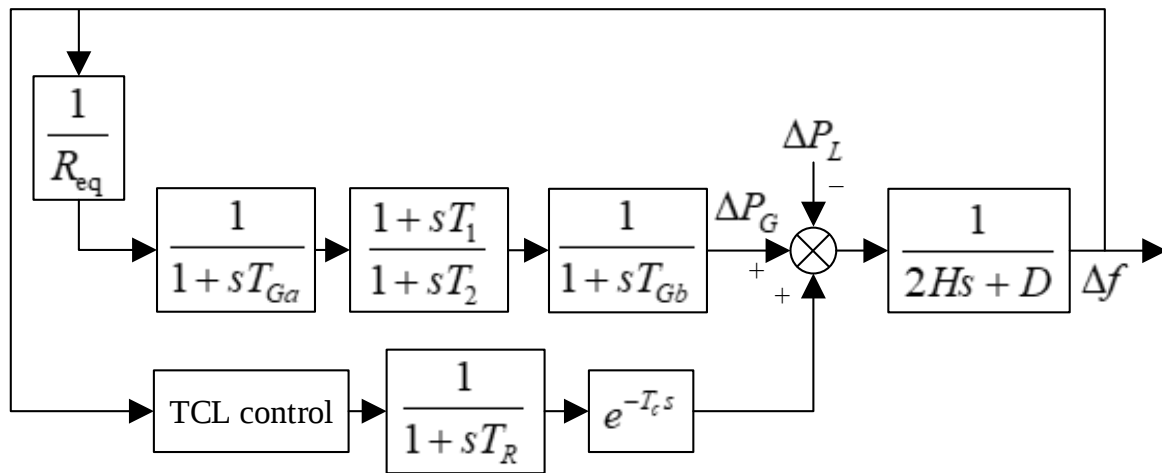


Figure 1. Frequency modulation model of power system with TCL participation.

2.2. TCL Control Methods

The control methods of TCL are mainly divided into direct switch control and temperature setting control. The TCL operation characteristics of the direct switch are shown in Figure 2a. The temperature setting value of the load remains unchanged, and the dispatching command directly acts on the equipment switch during the operation time $k = k_0$. The advantage of this method is that it can accurately track the power within the adjustable temperature range and has little impact on the user's comfort. However, when the indoor temperature approaches the temperature boundary, it will cause frequent switching, thus reducing the service life of the equipment. The TCL operating characteristics of the temperature setting are shown in Figure 2b. The dispatching command increases the temperature setting value at time $k = k_0$. Since the temperature dead zone of the load is unchanged, its operating range will change, thus indirectly changing the switching state of the equipment. Since the indoor temperature of the load is always uniformly distributed in the temperature dead zone, the indoor temperature will not approach the temperature boundary, but it has a notable impact on the user's comfort, and its tracking effect depends on the designed controller.

In practical application, the reasonable distribution of power regulation can make load clusters with different control modes cooperate with each other, which can not only realize accurate tracking of power, but also avoid their limitations and meet the different needs of users. Many factors need to be considered when allocating the power regulation amount. On the one hand, it is necessary to meet the requirements of the power system for frequency regulation and ensure the tracking accuracy within a certain range. On the other hand, for the load with direct switch, the number of switches should be reduced as much as possible to prolong the service life of the equipment. For the load with temperature setting, the temperature change should be reduced as much as possible to improve the user's comfort. There is often a highly nonlinear relationship between these factors and power distribution, and the frequency regulation requires high real-time performance. Therefore, the conventional optimization method cannot obtain the optimal power allocation.

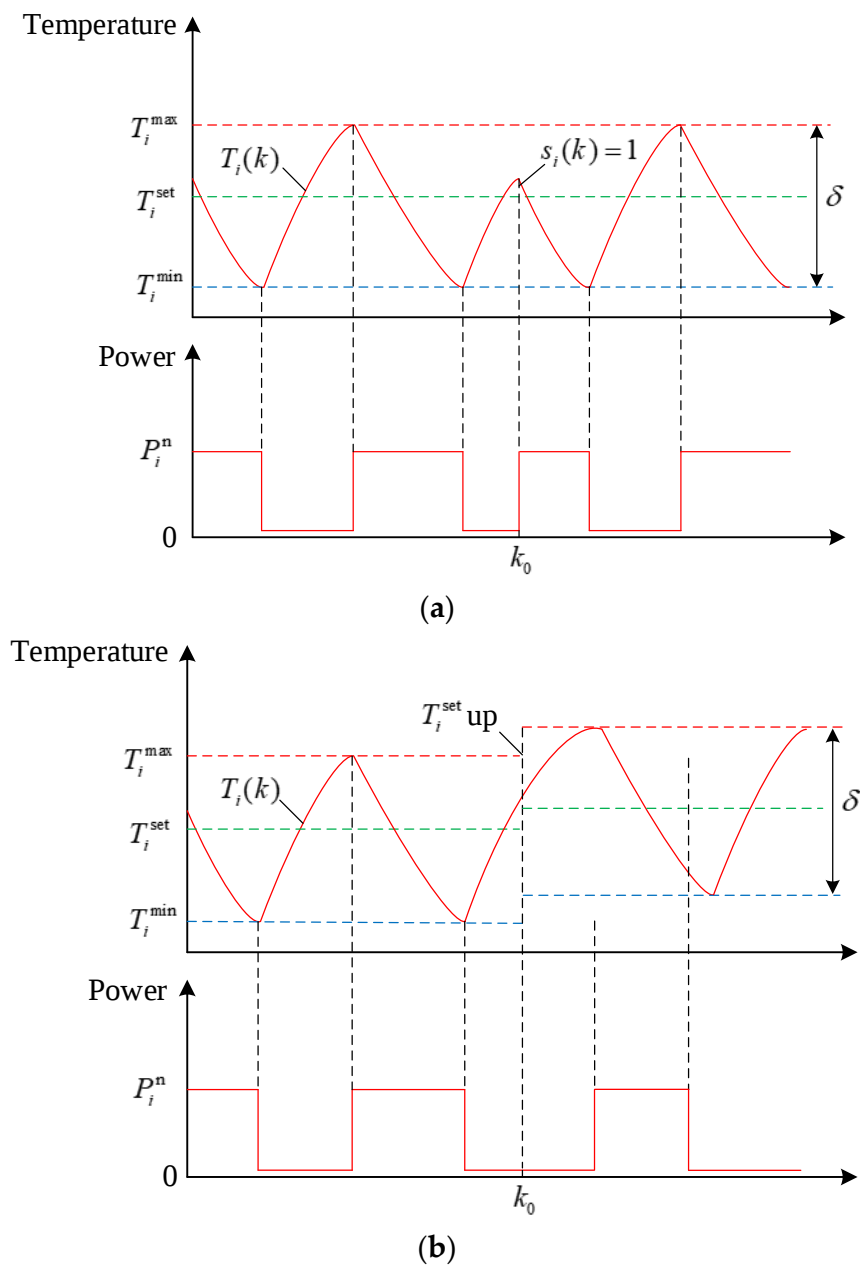


Figure 2. Operation characteristics of TCL in different control modes. (a) Direct switch control; (b) temperature setting control.

3. Comprehensive Control of TCL Considering Customer Satisfaction

3.1. Calculation of TCL Load Regulation Index

The temperature setting value of the direct switch load is almost constant, and the influence on the user's comfort is negligible. Its control mode directly acts on the equipment switch. When the indoor temperature is close to the temperature boundary, the load regulation capacity will decrease, and the start and stop frequency of the equipment will be higher and the service life will be shortened. When the indoor temperature is close to the temperature setting value, the load regulation capacity will rise. At this time, the start and stop frequency of the equipment is low and the service life is extended. In order to characterize the regulation ability of TCL and provide a basis for evaluating the service life

of equipment, we use the definition of battery state of charge for reference, and define the energy storage index C_s for TCL cluster with a direct switch under refrigeration mode:

$$C_s = \frac{2}{\delta} \left| \frac{\sum_{i=1}^N T_{\max} - T_i}{N} - \frac{\delta}{2} \right| \quad T_i \in (T_i^{\min}, T_i^{\max}) \quad (7)$$

According to the definition of C_s , the closer C_s is to 0, the closer the indoor temperature is to the temperature set value. At this time, the temperature distribution of TCL is relatively uniform, its adjustable potential is large and the switch switching is not frequent. The closer C_s is to 1, the closer the indoor temperature is to the upper and lower temperature limits. At this time, the temperature distribution of TCL is more concentrated, its adjustable potential is small and the switch switching is more frequent. Therefore, when the agent outputs the control command, it should make C_s close to 0 as much as possible, so as to reduce the start and stop frequency of the equipment.

The tracking effect of the load depends on the designed controller, and because the set value changes, the temperature changes in a large range, which will reduce the comfort level of the user. To characterize the user's comfort, the discomfort index C_u is defined as follows for TCL cluster with temperature setting:

$$C_u = |T_i^{\text{set}} - T_i^{\text{set}}(0)| \quad (8)$$

where $T_i^{\text{set}}(0)$ indicates the initial temperature setting value. According to the definition of C_u , the more the temperature setting value deviates from the initial temperature setting value, the higher the user's discomfort level. Therefore, when the agent outputs the control command, it should make C_u as close to 0 as possible to reduce the user's discomfort.

3.2. Customer Satisfaction Assessment Method Based on Regulation Index

According to the above analysis, users' needs are different under different control methods. In order to comprehensively evaluate customer satisfaction, the fuzzy comprehensive evaluation (FCE) method is adopted for evaluation. The specific operations are as follows.

- (1) Construct the factor set of customer satisfaction $U = \{C_s, C_u\}$.
- (2) Build a customer satisfaction evaluation set $V = \{\text{satisfied, more satisfied, general, less satisfied, dissatisfied}\}$.
- (3) Determine the weight of each factor. Since the factor set in this paper is composed of two factors, C_s and C_u , which have the same importance to users, the weight is $A = [a_1, a_2] = [0.5, 0.5]$.
- (4) The fuzzy evaluation matrix is established. Firstly, the degree to which each factor belongs to each comment is evaluated. Since most things follow a normal distribution, the membership function is selected as a Gaussian function.

$$r_{sp}(y_s) = e^{-\left(\frac{y_s - u_{sp}}{\sigma_{sp}}\right)^2} \quad (9)$$

where y_s is the input of the s -th factor, C_s and C_u . u_{sp} and σ_{sp} are the mean and standard deviation of the s -th factor and the p -th comment, respectively. Then, the fuzzy evaluation matrix R is:

$$R = \begin{bmatrix} r_{11} & r_{12} & r_{13} & r_{14} & r_{15} \\ r_{21} & r_{22} & r_{23} & r_{24} & r_{25} \end{bmatrix} \quad (10)$$

- (5) Fuzzy comprehensive evaluation is carried out. The fuzzy evaluation set is:

$$B = A \circ R = [b_1 \ b_2 \ b_3 \ b_4 \ b_5] \quad (11)$$

where $\circ$ represents the operation of the fuzzy matrix. Since the weighted average fuzzy synthesis operator has an obvious weight effect and strong comprehensive degree, and can make full use of the information of R , the element b_p is:

$$b_p = \min \left(1, \sum_{s=1}^2 a_s r_{sp} \right) \quad (12)$$

- (6) Evaluate customer satisfaction. In order to make the level continuous and quantitative, the level rank corresponding to the matrix B element is set as 1, 2, 3, 4 and 5, and the customer satisfaction m is defined as:

$$m = \frac{b_1 + 2b_2 + 3b_3 + 4b_4 + 5b_5}{b_1 + b_2 + b_3 + b_4 + b_5} \quad (13)$$

According to the definition of m , the smaller m is, the higher the user's satisfaction.

In practical application, the agent should not only consider customer satisfaction, but also meet the requirements of the power system for frequency regulation, that is, to ensure the tracking accuracy within a certain range. In order to evaluate the tracking performance of the system, the root mean square error index E_{RMS} is defined as:

$$E_{RMS} = \sqrt{\frac{\sum_{\Delta k=1}^{N_s} (e(\Delta k))^2}{N_s (P_{\text{target}}^{\max} - P_{\text{target}}^{\min})^2}} \times 100\% \quad (14)$$

where N_s is the number of control cycle Δk , $e(\Delta k)$ is an error signal in a control period, and $P_{\text{target}}^{\max}$ and $P_{\text{target}}^{\min}$ are the minimum value and the maximum value of the tracking power signal, respectively.

According to the definition of E_{RMS} , the smaller E_{RMS} is, the higher the tracking accuracy of the system. In order to comprehensively evaluate the regulation effect and provide the basis for the optimization of the power distribution signal, the comprehensive evaluation index J is defined as:

$$J = (1 - \lambda)E_{RMS} + \lambda m \quad (15)$$

where λ is the proportion of satisfaction.

In fact, priority should be given to ensuring the frequency stability of the power grid: when the tracking accuracy is within a certain range, customer satisfaction can be considered; otherwise, it will not be considered. The relationship between λ and E_{RMS} is as follows:

$$\lambda = \begin{cases} G_1 & 0 < E_{RMS} \leq F_1 \\ G_2 & F_1 < E_{RMS} \leq F_2 \\ G_3 & F_2 < E_{RMS} \leq F_3 \\ 0 & E_{RMS} > F_3 \end{cases} \quad (16)$$

where F_1, F_2, F_3, G_1, G_2 and G_3 are constants, which can be set according to actual operation conditions.

4. Frequency Response Control of TCL Based on SAC Deep Reinforcement Learning

4.1. A Deep Reinforcement Learning Model of Soft Actor-Critic

Reinforcement learning is adaptive learning in the way of trial and error of agents. The agent interacts with the environment continuously, and takes actions to change the environmental state by acquiring the environmental state. The agent obtains corresponding rewards or punishments as the update guidance of the model parameters, so as to obtain the maximum cumulative rewards in continuous learning. Through this perception action evaluation learning method, the agent continuously obtains knowledge in the interaction

process, and constantly adjusts and improves its action strategy to adapt to the environment, and finally gives a better task execution strategy. The environment interaction mode is generally described by the Markov decision process (MDP) and is composed of five tuples (S, A, P, r, γ) , namely state space S , action space A , state transition probability P , return function r and discount factor γ .

In this paper, the deep reinforcement learning based on the flexible actor evaluator framework is used to control the frequency response of TCL. The framework of the proposed control model is shown in Figure 3. In the iterative calculation at time t , the actor first generates the action a_t through the policy network according to the operating state s_t of the TCL cluster observed at this time. After that, the TCL cluster performs state transition according to the control strategy at this time, and reaches the state s_{t+1} at the next time. At the same time, the system environment calculates the reward $r(s_t, a_t)$ at time t and feeds it back to the agent, who records $(s_t, a_t, r(s_t, a_t), s_{t+1})$ in the experience pool. Then, the action strategy sampling of the actor and the system state are input to the critic at the same time, and the action value function $Q(s_t, a_t)$ is output to evaluate the strategy. This process is carried out circularly, and the actor and the judge update their neural network parameters through the gradient descent method, so as to realize the adaptive learning of the model. During the training process, the accumulated return of the agent in the response period will gradually increase and eventually become stable. By introducing the maximum entropy encouragement strategy, the SAC reinforcement learning algorithm can improve the robustness of the algorithm and speed up the training speed. It can make accurate and effective control decisions for large-scale temperature control loads in the complex power supply and demand environment.

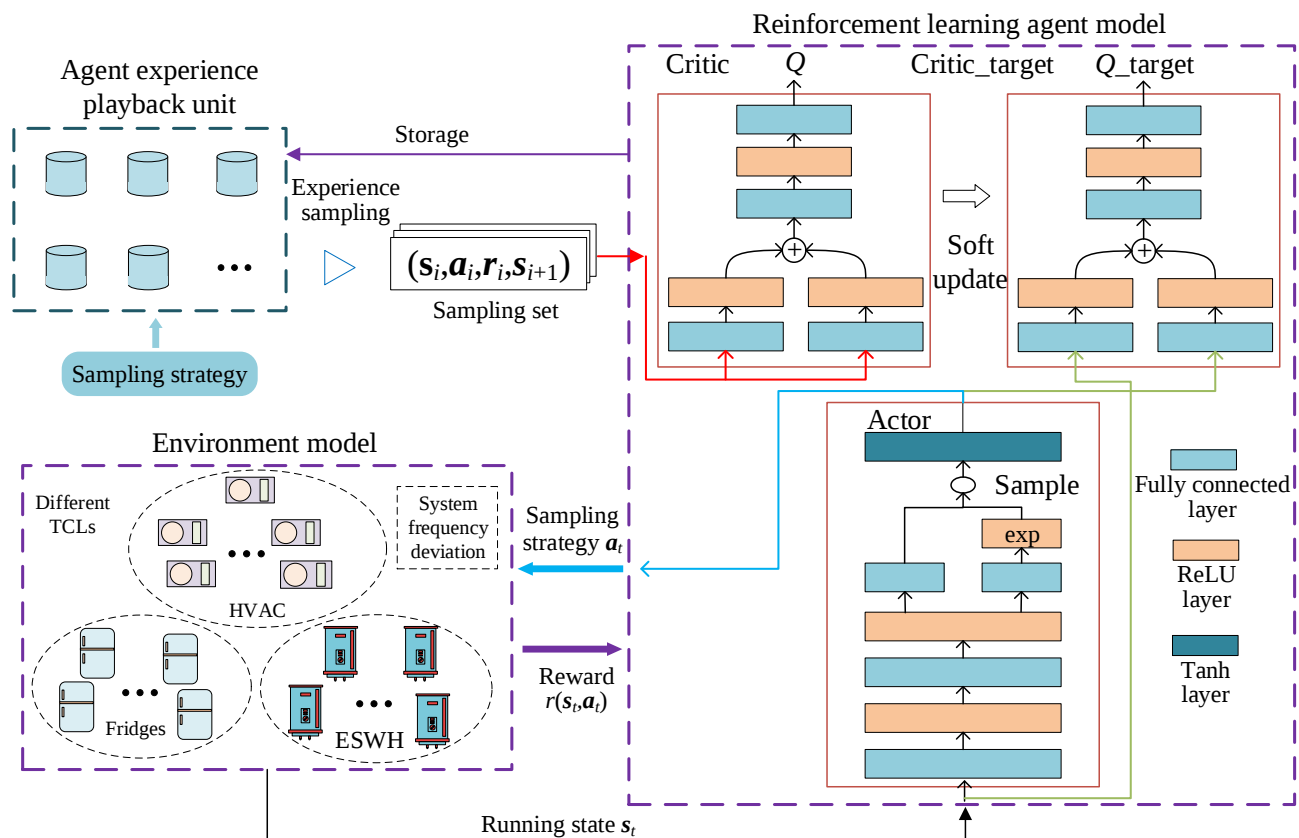


Figure 3. Frequency response control framework of large-scale TCLs based on deep reinforcement learning.

4.2. SAC Deep Reinforcement Learning Method

4.2.1. SAC Objective Function

The objective function of SAC requires the strategy to maximize the policy entropy while maximizing the cumulative return, so as to avoid greedy sampling in the learning process and falling into local optimization. According to this, the objective function $\pi_{\max}^*$ is constructed as shown in the equation.

$$\pi_{\max}^* = \arg \max_{\pi} \sum_{t=1}^T E_{(s_q, a_q) \sim p_{\pi}} (r(s_q, a_q) + \alpha H(\pi(\cdot|s_q))) \quad (17)$$

where $E(\cdot)$ is the expected function, π is a policy, s_q is the state space of the q -th agent, a_q is the action space of the TCL and $r(s_q, a_q)$ is the reward function of the q -th agent. The state action trajectory $(s_q, a_q) \sim p_{\pi}$ formed by the strategy π . α is the temperature term, which determines the influence of entropy on reward. $H(\pi(\cdot|s_q))$ is the entropy term of the strategy in the state, and its calculation method is shown in Equation (18).

$$H(\pi(a_q|s_q)) = - \int_{a_q} \pi(a_q|s_q) \log(\pi(a_q|s_q)) da_q = E_{a_q \sim p_{\pi}} (-\log(\pi(a_q|s_q))) \quad (18)$$

4.2.2. SAC Iteration Strategy

The value function in the reinforcement learning process is shown in Equation (19), which is used for strategic value evaluation $Q(s_q, a_q)$. The bellman backup operator is shown in Equation (20) and is used for policy updating.

$$Q(s_q, a_q) = r(s_q, a_q) + \gamma E_{s_{t+1} \sim p} (Q(s_{q+1}, a_{q+1})) \quad (19)$$

$$T^{\pi} Q(s_q, a_q) \triangleq r(s_q, a_q) + \gamma E_{s_{t+1} \sim p} (V(s_{q+1})) \quad (20)$$

where $E_{s_{t+1} \sim p}$ is the expected function of the state space, T^{π} is the bellman backup operator under the policy π and γ is the discount factor of the reward. $V(s_{q+1})$ is a new value function of the state, and its calculation method is shown in the following equation.

$$V(s_q) = E_{a_q \sim \pi} (Q(s_q, a_q) - \log \pi(a_q|s_{q+1})) \quad (21)$$

Meanwhile, there are

$$Q^{k+1} = T^{\pi} Q^k \quad (22)$$

where Q^k is the value function of the k th calculation.

Equation (23) can be realized by iterating Equations (20) and (22) continuously.

$$\lim_{k \rightarrow \infty} Q^k = \hat{Q} \quad (23)$$

where $\hat{Q}$ is the soft Q -value.

4.2.3. SAC Policy Update

The strategy updating method in the calculation process is shown in Equation (24).

$$\pi_{new} = \arg \min_{\pi \in \Pi} D_{KL} \left(\pi(\cdot|s_q) \parallel \frac{\exp\left(\frac{1}{\alpha} Q^{\pi_{old}}(s_q, \cdot)\right)}{Z^{\pi_{old}}(s_q)} \right) \quad (24)$$

where KL is divergence, Π is a policy set and $Q^{\pi_{old}}(\cdot)$ is the value function under the old strategy π_{old} . $Z^{\pi_{old}}(s_q)$ is the partition function under the old strategy π , which is used to normalize the distribution.

4.2.4. Construction of SAC Algorithm

The SAC algorithm needs to construct neural networks, including Q -value network and strategy network. The Q -value network outputs a single value through several layers of neural networks, and the strategy network outputs a Gaussian distribution. In this process, the neural network will be updated. The Q -value network parameter has an update strategy as shown in Equation (25), and the policy network parameter has an update strategy as shown in Equation (26).

$$J_Q(\theta) = E_{(s_q, a_q, s_{q+1}) \sim D} \left(\frac{1}{2} (Q(s_q, a_q) - (r(s_q, a_q) + \gamma V_{\bar{\theta}}(s_{q+1})))^2 \right) \quad (25)$$

$$J_\pi(\phi) = D_{KL} \left(\pi(\cdot | s_q) || \exp \left(\frac{1}{\alpha} Q_\theta(s_q, \cdot) - \log Z(s_q) \right) \right) \quad (26)$$

where θ is the network parameter of Q -value, ϕ is the policy network parameter and $V_{\bar{\theta}}$ and Q_θ are the new value function and the value function after substituting the Q -value network parameters. $Z(s_q)$ is the partition function of the state.

The temperature parameter is an important parameter to assist in maximizing entropy, which can maximize the exploration of action space. A reasonable temperature parameter setting is helpful to realize iterative testing of all feasible actions. Therefore, the update of the temperature parameter is as shown in Equation (27).

$$J(\alpha) = E_{a_q \sim \pi_q} (-\alpha \log \pi_q(a_q | \pi_q) - \alpha H_0) \quad (27)$$

where π_q is the control strategy of the q -th agent. H_0 is the entropy term.

The expressions (25)–(27) are all obtained by calculation. Throughout the whole process, the Q -value network parameters, strategy network parameters and temperature parameters are continuously updated through deep neural network learning, which can make the model converge continuously and solve the optimal strategy.

4.3. Design of SAC Deep Reinforcement Learning Model for TCL Frequency Response

In this paper, the SAC algorithm is used to solve the control strategy problem of large-scale temperature control load participating in system frequency regulation. The structural model of the proposed control method is shown in Figure 3. The agent in the figure is an agent based on a deep neural network. The environment of the controller is the frequency Δf deviation of the power system, the differentiation and integration of Δf , the baseline power signal P_{base} of the aggregated TCL, the aggregated power consumption P_{total} and the automatic generation control signal P_{AGC} . The automatic generation control signal is a series of positive and negative power signals, which represents the active power deviation between the supply and demand of the power system. The baseline power signal of aggregated TCL is the sum of the rated power $P_{i.set}^n$ of TCL cluster under a certain temperature setting value.

$$P_{base} = \sum_{i=1}^N P_{i.set}^n \quad (28)$$

$$P_{i.set}^n = \frac{T_\infty - T_i^{set}}{\eta_i R_i} \quad (29)$$

A tracking power signal P_{target} can be generated by superimposing the P_{AGC} and P_{base} . The tracking error signal e can be generated by subtracting the P_{target} and P_{total} , that is

$$P_{target} = P_{AGC} + P_{base} \quad (30)$$

$$e = P_{target} - P_{total} \quad (31)$$

The agent obtains the response power of TCL cluster after optimization according to the environmental information. Then, the direct switch control cluster and the temperature

setting control cluster complete the adjustment task according to the obtained control signals. The method includes two stages: offline pre-learning and online application. In the offline pre-learning stage, the pre-learning process will iteratively update all parameters of the agent. During each self-learning iteration, the agent will conduct action exploration (i.e., generate different commands) to interact with the environment. After exploration, the parameters of the agent will be updated according to the system frequency deviation and the reward function of the TCL controller. With an appropriate reward function R and considering environmental constraints, the gradient of the actor (i.e., the gradient of the control target relative to the parameters of the agent) will be calculated and used to update all the parameters of the agent. In the online application stage, the intelligent personnel will calculate the operation value (i.e., generate command) according to the observation value and parameters obtained by themselves for each control cluster.

For the frequency response model of large-scale temperature control load, the negative value of the comprehensive evaluation index J is taken as the reward function of the agent, that is

$$R = J = (1 - \lambda)E_{\text{RMS}} + \lambda m \quad (32)$$

By introducing the system frequency deviation and customer satisfaction into the reward function, the obtained control strategy can improve the tracking accuracy of the system, reduce the switching frequency of the equipment, reduce the temperature change and improve the customer satisfaction.

5. Result Analysis

5.1. Example Introduction and Scenario Setting

In order to verify the effectiveness of the method proposed in this paper for large-scale thermostatically controlled loads to participate in power grid frequency control, we take the distribution network with large-scale thermostatically controlled loads as an example, and set the disturbance power of the regional power grid as the net load power, that is, the difference between the original load power and the power generated by new energy, such as photovoltaic and wind power. The disturbance power in the simulation time is shown in Figure 4. As an important thermostatically controlled load, HVAC accounts for a large proportion and is easy to control and manage. Therefore, we selected 2000 HVAC units for simulation experiments. The load parameter settings are shown in Table 1. The initial indoor temperature of the load is evenly distributed in the temperature dead zone, and the temperature dead zone is set to 1.2 °C. P_{ACC} is the actual frequency regulation signal from the PJM power market in the United States, which changes every four seconds.

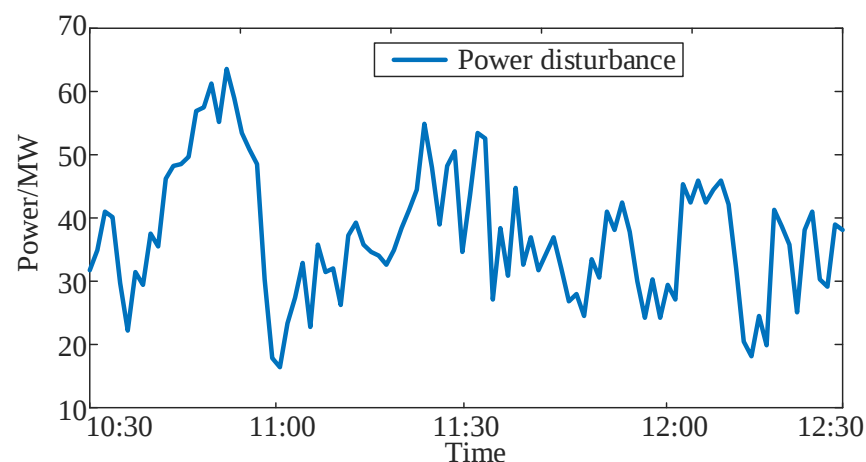


Figure 4. Variation of disturbance power within 2 h.

Table 1. Setting of load parameters.

Parameters	Meaning	Value
C_i	Equivalent heat capacity	$N(2, 0.01)$
R_i	Equivalent thermal resistance	$N(2, 0.01)$
P_i	Energy transfer rate	$N(14, 0.01)$
T_∞	Outdoor temperature	32
$T_i^{set}(0)$	Initial temperature setting	20
δ	Temperature dead zone	1.2
η_i	Energy conversion efficiency coefficient	2.5

Note: $N(2, 0.01)$ represents the normal distribution with the mean value of 2 and the variance of 0.01, and the others are the same.

The parameter settings in the customer satisfaction evaluation and comprehensive evaluation indicators are shown in Table 2. When the value of $ERMS$ is $(0, 2\%]$, the frequency regulation effect of the power system is better. At this time, customer satisfaction is mainly considered. When the value of E_{RMS} is $(2, 3\%]$, λ is 0.5, indicating that the root mean square error index and the customer satisfaction have the same impact on the system regulation. When the value of E_{RMS} is $(3\%, 5\%]$, the frequency regulation effect of the power system is poor, and the root mean square error index is mainly considered at this time. When E_{RMS} is greater than 5%, λ is 0, indicating that the system regulation will no longer consider the customer satisfaction, and will focus on improving the tracking accuracy.

Table 2. Setting of customer satisfaction evaluation and comprehensive evaluation parameters.

Parameters	Value	Parameters	Value
u_{s1}	0	F_1	2%
u_{s2}	0.25	F_2	3%
u_{s3}	0.5	F_3	5%
u_{s4}	0.75	G_1	0.8
u_{s5}	1	G_2	0.5
σ_{sp}	0.2	G_3	0.3

5.2. Frequency Control Effect Analysis Considering Customer Satisfaction

Figure 5 compares the changes in customer satisfaction before and after customer satisfaction is included in the agent optimization process. As can be seen from Figure 5, on the one hand, the peak value of customer satisfaction increased from 3.5 before optimization to 4.7 after optimization. On the other hand, within the simulation time range, customer satisfaction after optimization is higher than that before optimization. Figure 6 compares the frequency control effect of the method proposed in this paper before and after considering customer satisfaction. As can be seen from Figure 6, since the customer satisfaction index is added to the control target, the algorithm's penalty weight for frequency deviation is relatively reduced, resulting in an increase in the maximum frequency deviation of the system compared with that before considering customer satisfaction. Considering customer satisfaction will limit the number of TCLs participating in frequency response, resulting in a poor system frequency control effect, but the impact is not significant. Therefore, the method proposed in this paper can better balance the frequency deviation control and customer satisfaction.

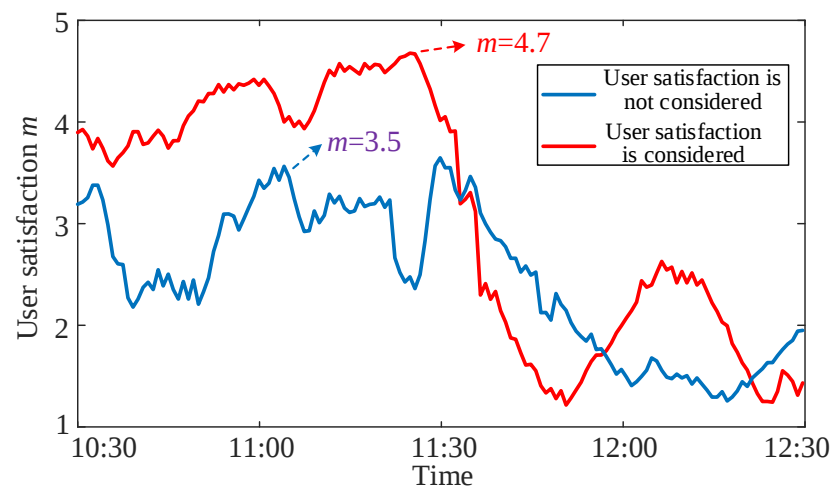


Figure 5. Comparison of customer satisfaction before and after optimization.

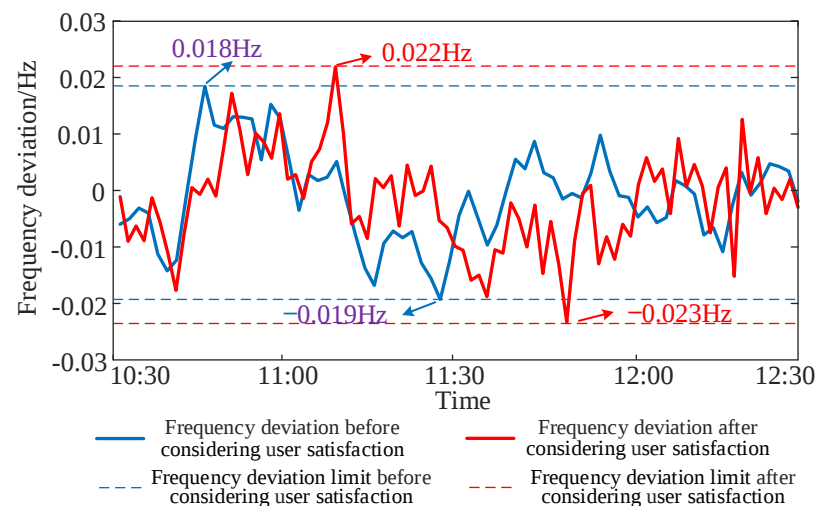


Figure 6. Frequency deviation before and after considering customer satisfaction.

5.3. Frequency Control EFFECT Analysis Based on SAC Deep Reinforcement Learning

In order to verify the effectiveness of the proposed SAC deep reinforcement learning algorithm in the collaborative control of large-scale thermostatically controlled load compared with the traditional PID method, we used the traditional PID controller, the PID controller optimized by particle swarm optimization (PSO) algorithm parameters, and the algorithm proposed in this paper to conduct simulation experiments on the thermostatically controlled load controlled by the direct switch control and the temperature setting control. The number of HVAC experiments is 2000. The comparison of the three control methods in the system frequency control effect is shown in Figure 7.

It can be seen from Figure 7 that under the regulation control of the conventional PID controller, there is a system frequency deviation of about 0.047 Hz in the period when the disturbance power fluctuates violently. Compared with the traditional PID controller, the frequency effect of the PID controller is improved after PSO algorithm parameter optimization, but the effect is still not ideal. The frequency deviation of the system fluctuates within the range of (−0.037 Hz, 0.038 Hz). The algorithm proposed in this paper can keep the system frequency deviation within (−0.02 Hz, 0.023 Hz), and can significantly improve the frequency stability of the power grid.

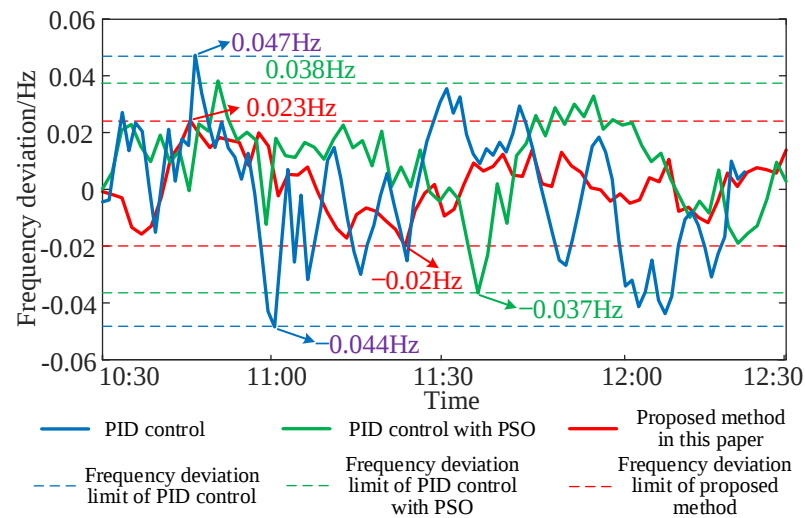


Figure 7. Comparison of frequency control effect between the proposed algorithm and PID controller.

5.4. Comparative Analysis of the Algorithm and Traditional Deep Reinforcement Learning

This section shows the advantages of the algorithm in this paper compared with the data-driven deep Q network (DQN) algorithm and the DDPG algorithm from the aspects of system frequency control effect and algorithm convergence speed. The proposed algorithm and DDPG are both deep reinforcement learning algorithms based on continuous action space. As shown in Figure 8, the frequency control effect of the two algorithms is significantly better than that of the DQN algorithm designed based on the discrete action space. After integrating the advantages of model driven and data driven, the algorithm proposed in this paper further improves the real-time frequency control effect in the continuous action domain compared with the fully trained DDPG algorithm.

Figure 9 compares the iterative convergence process of the cumulative reward value of the DQN algorithm, the DDPG algorithm and the algorithm proposed in this paper. Among them, after about 250 and 300 iteration cycles, the cumulative reward value of DQN algorithm and DDPG algorithm tends to be stable and will not continue to increase. It is worth noting that each iteration cycle in this paper is an empirical trajectory containing 200 iterations. That is, the two algorithms need 50,000 and 60,000 iterations, respectively, to converge. However, the proposed algorithm only needs about 150 iteration cycles (30,000 iterations) to complete the parameter training of the deep neural network, and the oscillation amplitude of the convergence curve of the algorithm is the smallest.

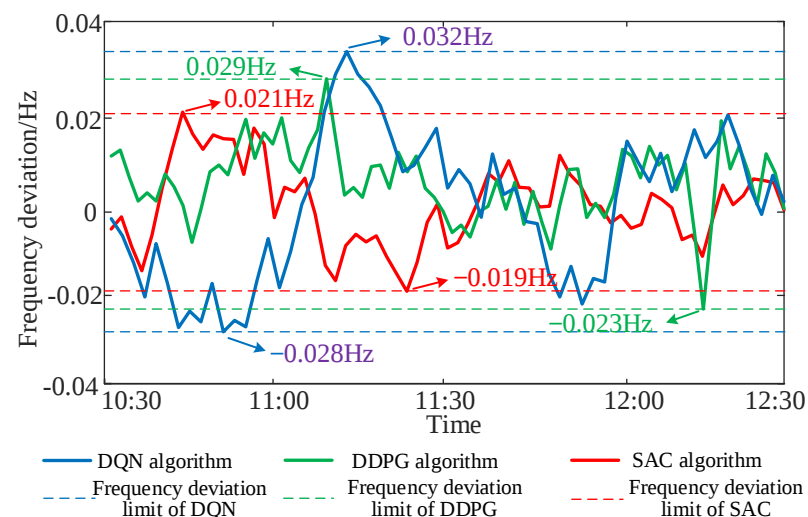


Figure 8. Comparison of frequency control effect between different algorithms.

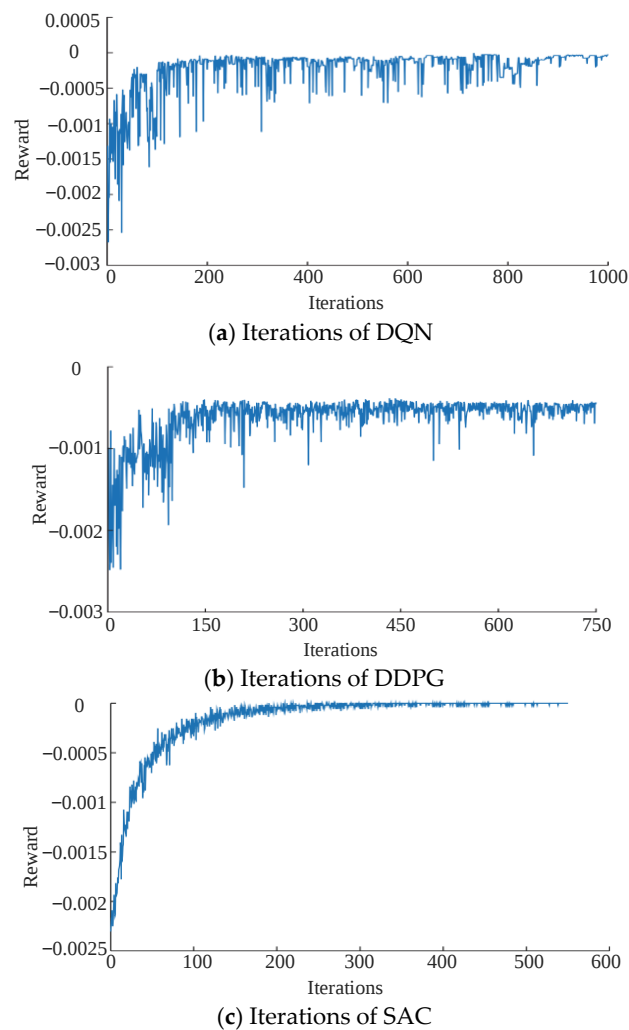


Figure 9. Iterative convergence process of cumulative reward value: (a) DQN algorithm; (b) DDPG algorithm; (c) SAC algorithm in this paper.

6. Conclusions

In this paper, considering the support of large-scale thermostatically controlled load on the demand side to the frequency of power grid, based on the deep reinforcement learning of soft actor–critic, a frequency cooperative control method of thermostatically controlled load considering customer satisfaction is proposed to solve the frequency control problem of a power system with large-scale thermostatically controlled load on the demand side participating in frequency regulation. In the example analysis, a distribution network is taken as the research object, and the performance of different algorithms is compared and verified based on time domain simulation. The simulation results show that compared with the existing deep reinforcement learning methods, this algorithm has obvious advantages in system frequency control, customer satisfaction and algorithm training time.

The algorithm proposed in this paper is mainly used to solve the frequency cooperative control problem under the participation of large-scale thermostatically controlled load in the distribution network, and does not consider the application of other flexible resources on the demand side in the system frequency modulation. The next step is to explore the application and practice of multi-agent deep reinforcement learning method in the frequency response of demand-side flexible resources based on the operation mechanism and mathematical modeling of demand-side flexible resources.

Author Contributions: Conceptualization, R.C. and G.Y.; Data curation, C.L. and X.Y.; Investigation, H.L. and Y.Z.; Methodology, R.C.; Software, H.L.; Supervision, G.Y. and Y.Z.; Validation, G.Y.; Visualization, X.Y.; Writing—original draft, R.C. and H.L.; Writing—review and editing, C.L., X.Y. and Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Science and Technology Project of State Grid Hubei Electric Power Research Institute: “Research on frequency dynamic prediction and active control strategy of high proportion new energy power system under mutational weather” (project number B31532225680).

Data Availability Statement: The data in this paper are from a real distribution network and involve a confidentiality agreement. The dataset in this paper is not publicly available.

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

Abbreviation

TCL	thermostatically controlled load
ESWH	heat pump, electric storage water heater
HVAC	heating, ventilation and air conditioning
LA	load aggregator
EMS	energy management system
DDPG	deep deterministic policy gradient
SAC	soft actor–critic
DQN	Deep Q network
DSAC	distributed soft actor–critic
FCE	fuzzy comprehensive evaluation
MDP	Markov decision process
PSO	particle swarm optimization
T_i	The indoor temperature of the i -th TCL
T_∞	The outdoor temperature
C_i	The equivalent heat capacity of the i -th TCL
R_i	The equivalent heat resistance of the i -th TCL
P_i	The energy transfer rate of the i -th TCL
$s_i(k)$	The load switch state of TCL
$T_i^{\max}$	The upper limits of the temperature during load operation
$T_i^{\min}$	The lower limits of the temperature during load operation
T_i^{set}	the temperature setting value
δ	the temperature dead zone interval
P_i^n	the rated power of the i -th TCL
η_i	the energy conversion efficiency coefficient
C_s	the energy storage index for TCL cluster with direct switch control
C_u	the discomfort index for TCL cluster with temperature setting control
r_{sp}	the membership function of the s -th factor and the p -th comment
y_s	the input of the s -th factor
u_{sp}	the mean of the s -th factor and the p -th comment
σ_{sp}	the standard deviation of the s -th factor and the p -th comment
b_p	the element of fuzzy evaluation set B
m	the customer satisfaction index
E_{RMS}	the root mean square error index of frequency
$e(\Delta k)$	the error signal in a control period
$p_{\text{target}}^{\max}$	the maximum value of the tracking power signal
$p_{\text{target}}^{\min}$	the minimum value of the tracking power signal
J	the comprehensive evaluation index
λ	the proportion of satisfaction
Δf	the frequency deviation of the power system
P_{base}	the baseline power signal of the aggregated TCLs
P_{total}	the aggregated power consumption of TCLs
P_{AGC}	the automatic generation control signal

References

- Li, C.; Wu, Y.; Sun, Y.; Zhang, H.; Liu, Y.; Liu, Y.; Terzija, V. Continuous Under-Frequency Load Shedding Scheme for Power System Adaptive Frequency Control. *IEEE Trans. Power Syst.* **2020**, *35*, 950–961. [\[CrossRef\]](#)
- Guan, M. Scheduled Power Control and Autonomous Energy Control of Grid-Connected Energy Storage System (ESS) with Virtual Synchronous Generator and Primary Frequency Regulation Capabilities. *IEEE Trans. Power Syst.* **2020**, *37*, 942–954. [\[CrossRef\]](#)
- Hosseini, S.A.; Toulabi, M.; Dobakhshari, A.S.; Ashouri-Zadeh, A.; Ranjbar, A.M. Delay Compensation of Demand Response and Adaptive Disturbance Rejection Applied to Power System Frequency Control. *IEEE Trans. Power Syst.* **2020**, *35*, 2037–2046. [\[CrossRef\]](#)
- Trovato, V.; Sanz, I.; Chaudhuri, B.; Strbac, G. Advanced control of thermostatic loads for rapid frequency response in Great Britain. *IEEE Trans. Power Syst.* **2017**, *32*, 2106–2117. [\[CrossRef\]](#)
- Bao, Y.; Li, Y.; Hong, Y.; Wang, B. Design of a hybrid hierarchical demand response control scheme for the frequency control. *IET Gener. Transm. Distrib.* **2015**, *9*, 2303–2310. [\[CrossRef\]](#)
- Mallada, E.; Zhao, C.; Low, S. Optimal load-side control for frequency regulation in smart grids. *IEEE Trans. Autom. Control* **2017**, *62*, 6294–6309. [\[CrossRef\]](#)
- Yu, Y.; Quan, L.; Mi, Z.; Lu, J.; Chang, S.; Yuan, Y. Improved Model Predictive Control with Prescribed Performance for Aggregated Thermostatically Controlled Loads. *J. Mod. Power Syst. Clean Energy* **2022**, *10*, 430–439. [\[CrossRef\]](#)
- Wang, J.; Shi, Y.; Zhou, Y. Intelligent demand response for industrial energy management considering thermostatically controlled loads and EVs. *IEEE Trans. Ind. Informatics* **2019**, *15*, 3432–3442. [\[CrossRef\]](#)
- Hu, J.; Cao, J.; Chen, M.Z.Q.; Yu, J.; Yao, J.; Yang, S.; Yong, T. Load Following of Multiple Heterogeneous TCL Aggregators by Centralized Control. *IEEE Trans. Power Syst.* **2017**, *32*, 3157–3167. [\[CrossRef\]](#)
- Delavari, A.; Kamwa, I. Improved Optimal Decentralized Load Modulation for Power System Primary Frequency Regulation. *IEEE Trans. Power Syst.* **2018**, *33*, 1013–1025. [\[CrossRef\]](#)
- Song, M.; Sun, W.; Wang, Y.; Shahidepour, M.; Li, Z.; Gao, C. Hierarchical Scheduling of Aggregated TCL Flexibility for Transactive Energy in Power Systems. *IEEE Trans. Smart Grid* **2020**, *11*, 2452–2463. [\[CrossRef\]](#)
- Wang, Y.; Tang, Y.; Xu, Y.; Xu, Y. A Distributed Control Scheme of Thermostatically Controlled Loads for the Building-Microgrid Community. *IEEE Trans. Sustain. Energy* **2020**, *11*, 350–360. [\[CrossRef\]](#)
- Kim, Y.J.; Fuentes, E.; Norford, L.K. Experimental Study of Grid Frequency Regulation Ancillary Service of a Variable Speed Heat Pump. *IEEE Trans. Power Syst.* **2016**, *31*, 3090–3099. [\[CrossRef\]](#)
- Song, M.; Gao, C.; Yan, H.; Yang, J. Thermal Battery Modeling of Inverter Air Conditioning for Demand Response. *IEEE Trans. Smart Grid* **2018**, *9*, 5522–5534. [\[CrossRef\]](#)
- Zhao, H.; Wu, Q.; Huang, S.; Zhang, H.; Liu, Y.; Xue, Y. Hierarchical Control of Thermostatically Controlled Loads for Primary Frequency Support. *IEEE Trans. Smart Grid* **2018**, *9*, 2986–2998. [\[CrossRef\]](#)
- Luo, F.; Dong, Z.; Meng, K.; Wen, J.; Wang, H.; Zhao, J. An operational planning framework for large-scale thermostatically controlled load dispatch. *IEEE Trans. Ind. Inform.* **2017**, *13*, 217–227. [\[CrossRef\]](#)
- Zhang, W.; Lian, J.; Chang, C.Y.; Kalsi, K. Aggregated Modeling and Control of Air Conditioning Loads for Demand Response. *IEEE Trans. Power Syst.* **2013**, *28*, 4655–4664. [\[CrossRef\]](#)
- Zhang, R.; Chu, X.; Zhang, W.; Liu, Y. Active Participation of Air Conditioners in Power System Frequency Control Considering Users' Thermal Comfort. *Energies* **2015**, *8*, 10818–10841. [\[CrossRef\]](#)
- Bashash, S.; Fathy, H.K. Modeling and Control of Aggregate Air Conditioning Loads for Robust Renewable Power Management. *IEEE Trans. Control Syst. Technol.* **2013**, *21*, 1318–1327. [\[CrossRef\]](#)
- Zhao, H.; Zhao, J.; Shu, T.; Pan, Z. Hybrid-Model-Based Deep Reinforcement Learning for Heating, Ventilation, and Air-Conditioning Control. *Front. Energy Res.* **2021**, *8*, 610518. [\[CrossRef\]](#)
- Chen, Z.; Shi, J.; Song, Z.; Yang, W.; Zhang, Z. Genetic Algorithm Based Temperature-Queuing Method for Aggregated IAC Load Control. *Energies* **2022**, *15*, 535. [\[CrossRef\]](#)
- Pallonetto, F.; de Rosa, M.; Milano, F.; Finn, D.P. Demand response algorithms for smart-grid ready residential buildings using machine learning models. *Appl. Energy* **2019**, *239*, 1265–1282. [\[CrossRef\]](#)
- Tindemans, S.H.; Strbac, G. Low-Complexity Decentralized Algorithm for Aggregate Load Control of Thermostatic Loads. *IEEE Trans. Ind. Appl.* **2021**, *57*, 987–998. [\[CrossRef\]](#)
- Ma, K.; Yuan, C.; Liu, Z.; Yang, J.; Guan, X. Hybrid control of aggregated thermostatically controlled loads: Step rule, parameter optimization, parallel and cascade structures. *IET Gener. Transm. Distrib.* **2016**, *10*, 4149–4157. [\[CrossRef\]](#)
- Xi, L.; Li, H.; Zhu, J.; Li, Y.; Wang, S. A Novel Automatic Generation Control Method Based on the Large-Scale Electric Vehicles and Wind Power Integration Into the Grid. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**. [\[CrossRef\]](#) [\[PubMed\]](#)
- Peirelinck, T.; Hermans, C.; Spiessens, F.; Deconinck, G. Domain Randomization for Demand Response of an Electric Water Heater. *IEEE Trans. Smart Grid* **2020**, *12*, 1370–1379. [\[CrossRef\]](#)
- Ruelens, F.; Claessens, B.J.; Vandael, S.; de Schutter, B.; Babuška, R.; Belmans, R. Residential demand response of thermostatically controlled loads using batch reinforcement learning. *IEEE Trans. Smart Grid* **2017**, *8*, 2149–2159. [\[CrossRef\]](#)

-
28. Liu, H.; Xu, F.; Fan, P.; Liu, L.; Wen, H.; Qiu, Y.; Ke, S.; Li, Y. Load Frequency Control Strategy of Island Microgrid with Flexible Resources Based on DQN. In Proceedings of the 2021 IEEE Sustainable Power and Energy Conference (iSPEC), Nanjing, China, 23–25 December 2021; pp. 632–637. [[CrossRef](#)]
 29. Yan, Z.; Xu, Y. A multi-agent deep reinforcement learning method for cooperative load frequency control of a multi-area power system. *IEEE Trans. Power Syst.* **2020**, *35*, 4599–4608. [[CrossRef](#)]
 30. Yu, L.; Xie, W.; Xie, D.; Zou, Y.; Zhang, D.; Sun, Z.; Zhang, L.; Zhang, Y.; Jiang, T. Deep Reinforcement Learning for Smart Home Energy Management. *IEEE Internet Things J.* **2019**, *7*, 2751–2762. [[CrossRef](#)]
 31. Xie, J.; Sun, W. Distributional Deep Reinforcement Learning-Based Emergency Frequency Control. *IEEE Trans. Power Syst.* **2021**, *37*, 2720–2730. [[CrossRef](#)]
 32. Yan, Z.; Xu, Y. Data-driven load frequency control for stochastic power systems: A deep reinforcement learning method with continuous action search. *IEEE Trans. Power Syst.* **2019**, *34*, 1653–1656. [[CrossRef](#)]
 33. Acharya, S.; Moursi, M.S.E.; Al-Hinai, A. Coordinated Frequency Control Strategy for an Islanded Microgrid with Demand Side Management Capability. *IEEE Trans. Energy Convers.* **2018**, *33*, 639–651. [[CrossRef](#)]