District Cooling System Control for Providing Operating Reserve Based on Safe Deep Reinforcement Learning

Peipei Yu, *Graduate Student Member, IEEE,* Hongcai Zhang, *Member, IEEE,* Yonghua Song, *Fellow, IEEE,* Hongxun Hui, *Member, IEEE,* and Ge Chen, *Graduate Student Member, IEEE*

***Abstract*—Heating, ventilation, and air conditioning (HVAC) systems are well proved to be capable to provide operating reserve for power systems. As a type of large-capacity and energy-efficient HVAC system (up to 100 MW), district cooling system (DCS) is emerging in modern cities and has huge potential to be regulated as a flexible load. However, strategically controlling a DCS to provide flexibility is challenging, because one DCS services multiple buildings with complex thermal dynamics and uncertain cooling demands. Improper control may lead to significant thermal discomfort and even deteriorate the power system's operation security. To address the above issues, we propose a model-free control strategy based on the deep reinforcement learning (DRL) without the requirement of accurate system model and uncertainty distribution. To avoid damaging "trial & error" actions that may violate the system's operation security during the training process, we further propose a safe layer combined to the DDPG to guarantee the satisfaction of critical constraints, forming a safe-DDPG scheme. Moreover, after providing operating reserve, DCS increases power and tries to recover all the buildings' temperature back to set values, which may cause an instantaneous peak-power rebound and bring a secondary negative impact on power systems. Therefore, we design a self-adaption reward function within the proposed safe-DDPG scheme to constrain the peak-power effectively. Numerical studies based on a realistic DCS demonstrate the effectiveness of the proposed methods.**

***Index Terms*—District cooling system, operating reserve, model-free control, safe deep reinforcement learning.**

I. Introduction

A. Background

THE increasing intermittent renewable energy resources bring more uncertainties to the generation-side, and scale up the demands for operating reserve services in power systems [1]. Traditionally, the service is majorly provided by thermal or gas generating units, which are carbon-intensive and are being phased out [2]. With the development of Internet of Things technologies, active control of demand-side resources has emerged as an alternative solution to provide operating reserve by curtailing or transferring power consumption [3]. The *heating, ventilation, and air conditioning* (HVAC) system is an ideal resource, because it can shift its power consumption flexibly while assuring the comfortable temperature by utilizing the building's inherent thermal inertia. Besides, HVACs have large regulation capacity as they account for over 40% of the total power consumption in modern cities [4].

Compared with a common household HVAC system, the *district cooling system* (DCS) is one type of HVAC with larger capacity and higher efficiency, and thus DCS is emerging and being developed in many cities [5]. As shown in Fig. 1, a DCS is composed of one energy station and some pipelines to produce chilled water for multiple buildings [6]. Generally, one DCS's capacity can be up to 100 MW, which is more than 10,000 times of a household HVAC. To fill this research gap, this study focuses on the real-time control of a DCS to provide operating reserve subject to the comfortable temperature constraint in each building. In most electricity markets, the start time for resources (i.e, DCS) to provide operating reserve is uncertain, while the time interval for operating reserve is fixed (e.g., 10 minutes in PJM [7], 15∼30 minutes in China [8]). As illustrated by the load curve in Fig. 1, there are two control stages for a DCS to provide operating reserve:

1) *In the power reduction stage*, the controller cuts down the DCS operating power following the instruction from the power system operator. In the meantime, it also tries to fulfill the temperature requests of heterogeneous buildings, when the cooling supply from DCS gets decreased as a result of power reduction.
2) *In the power recovery stage*, the DCS stops providing reserve and begins to restore all the buildings' indoor temperature back to set values by increasing its cooling supply. During this stage, the DCS shall recover its power consumption smoothly to avoid the peak-power rebound that may cause a secondary impact on the power system, which has just returned to the stable state.

The above two-stage control of DCS is quite challenging because of both the system complexity and cooling demand uncertainty, detailed as follows:

Complexity: To provide operating reserve services, an accurate thermal dynamic model of DCS should generally be developed to describe the relationship between the operating power and mass flow. However, this is challenging because the thermal dynamics of a DCS, including cooling power generation, transportation and consumption, is usually quite complex [5]. Conventional model-based control methods for

This paper is funded in part by the National Natural Science Foundation of China under Grant 52007200 , and in part by the Science and Technology Development Fund, Macau SAR (File no. SKL-IOTSC(UM)-2021-2023, and 0003/2020/AKP). (Corresponding author: *Hongcai Zhang*.)

P. Yu, H. Zhang, Y. Song, H. Hui, and G. Chen, are with the State Key Laboratory of Internet of Things for Smart City and Department of Electrical and Computer Engineering, University of Macau, Macao, 999078 China (email: hczhang@um.edu.mo).

This article has been accepted for publication in IEEE Transactions on Power Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2023.3237888

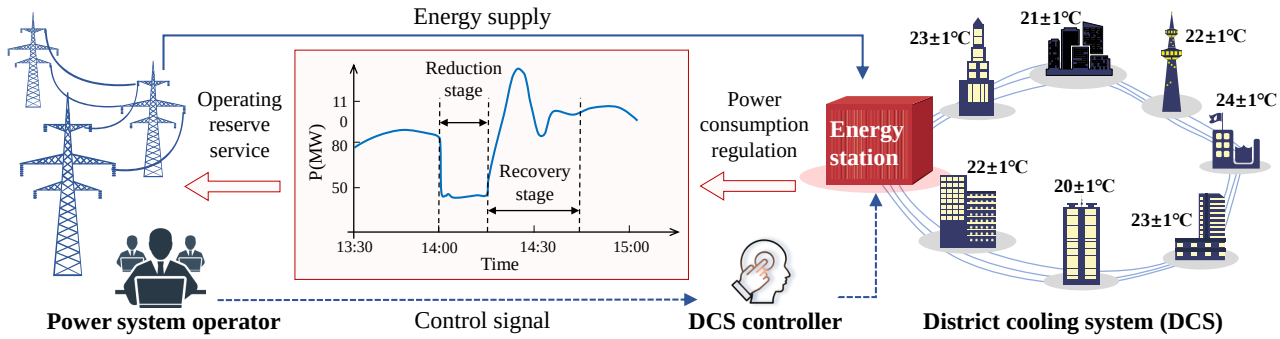


Fig. 1: The supply methods and scale of the DCS.

HVAC systems are hard to be used in DCS. Furthermore, the power consumption of a DCS is usually adjusted automatically according to its operation state, whose power cannot be controlled directly to provide operating reserve like other demand-side resources.

Uncertainty: The DCS's power consumption and buildings' indoor temperatures are related to the ambient temperature and indoor human behaviors. Higher ambient temperature and more indoor human activities call for more cooling supply and higher power consumption. However, the ambient temperature may bring different influences to heterogeneous buildings. The indoor human behaviors are also stochastic and hard to accurately predict [4]. As a result, it is nontrivial to control a DCS subject to heterogeneous indoor temperature constraints in multiple buildings, especially when the DCS power consumption is cut down to provide operating reserve.

B. Literature Reviews

In recent years, few published papers have studied the strategic operation of DCS to save electricity costs or provide flexible services to power grids. For example, Cox et al. [6] and Chen et al. [9] design day-ahead power scheduling strategies for DCS to minimize electricity costs with time-of-use pricing. Lo et al. [10] use least squares regression to optimize the day-ahead power dispatch for a large cooling system to perform demand response. Tang et al. [11] propose a direct load control strategy for a centralized AC system in response to requests of smart grid operators. The chillers are assumed to be operated in the on-off mode. The above studies focus on the day-ahead or hour-ahead operation of DCS while ignoring the real-time uncertainties in cooling demands. Besides, the operation mode of chillers is usually continuous so that assuming it to be on-off mode may not fully utilize DCS's regulation capacity.

Some other researchers have studied the real-time control of DCS. For example, the model predictive control (MPC) method is used to regulate DCS for achieving cost reduction in energy systems [12]. However, MPC usually requires a reliable dynamic model of the system and is often unavailable in practice, because it is quite computationally expensive and fails to work in real-time control scenarios. Mixed-integer linear programming (MILP) has also been used for the DCS power scheduling problems [13]. Unfortunately, at each step, optimization methods need to recalculate from the beginning, resulting in too a large calculation cost in real-time control. Moreover, the execution time of MILP increases exponentially according to the problem dimensions and cannot solve complex issues. Another category of commonly used control methods is heuristic algorithm, including genetic algorithm (GA), particle swarm optimization (PSO), and ants colony optimization (ACO), etc. For example, Stoppato et al. [14] combine heuristic algorithms to obtain system's optimal operation, while the convergence of heuristic algorithms cannot be proved mathematically and is less robust.

Compared with the aforementioned control methods, deep reinforcement learning (DRL) has become increasingly popular to handle model-free and high dimensional decision-making problems [15], such as building energy management [16]. DRL has been proved to be more robust with stable convergence results to effectively handle uncertainties of systems through the prediction in neural networks. Some researchers have adopted DRL to control traditional HVACs. For instance, Qiu et al. [17] control the HVAC to achieve energy saving based on DRL methods. Liu et al. [18] use the DRL method to find a near-optimal control strategy for exploiting the active and passive building thermal storage capacity. Du et al. [19] use DRL to control residential HVACs as to respond to dynamic electricity prices. Xu et al. [20] adopt DRL to schedule home energy consumption considering uncertain PV generation. Liang et al. [21] present a DRL-based control strategy to minimize both the HVAC's energy consumption and the user's thermal discomfort. However, to the best of our knowledge, published papers have not studied DRL-based control for DCS to provide the operating reserve for power systems.

Generally, a DRL-based controller has to be trained through lots of "trial-and-errors" before being intelligent [22]. It means some "bad" decisions may be made during training, part of which may cause critical constraint violations. However, in power systems, some critical constraint violations may cause damaging results [23]. For example, in the power reduction stage, if a DCS fails to provide sufficient operating reserve as it promised during services, the power system may face the risk of system collapse triggered [24]. Similarly, in the power reduction stage, if the DCS has a significant power rebound after providing services, the increased load current derived from the rebound peak may even harm system security considerably [25]. Therefore, the traditional approach is not an

ideal choice to cope with such critical constraints, because its policy has to learn along with frequent constraint violations.

To overcome the aforementioned limitation of conventional DRL algorithms, the safe-DRL framework is proposed to ensure the satisfaction of critical constraints during the training process. Comparing with conventional DRL algorithms that penalize constraints violations in the reward function, safe-DRL algorithms mainly adopt the following three methods to guarantee the satisfaction of critical constraints during training: 1) revising the agent's policy optimization rule considering critical constraints based on the constrained Markov decision process (CMDP), e.g., constrained policy optimization (CPO) [26], reward CPO [27], and projection-based CPO [28]; 2) limiting the agent's exploration process within the defined safe region, through the Gaussian estimation [29] or added safety networks [30]; 3) adding a safe layer to intervene in the agent's output actions, and tune the unsafe actions to safe ones through human or expert experiences [31]. The first two methods consider the long-term cost of constraints through expectations, but they cannot guarantee the satisfaction of the constraint at every time step. The third method can ensure zero constraint violations, but the design of the safe layer is challenging. To the best of our knowledge, there are no published papers that have adopted safe-DRL algorithms in HVAC or DCS control problems.

C. Contributions

In this paper, we propose a safe-DRL control strategy for DCS to provide operating reserve while satisfying major critical constraints. To the best of our knowledge, this is the first paper that has studied the DCS control problem for providing operating reserves and the first time a safe-DRL algorithm is used for DCS control. In summary, this paper advances the published literature in the following aspects:

1) The DCS control problem is developed as a Markov Decision Process (MDP) mathematically to provide operating reserves, which considers both the service performance and temperature comfort. Besides, DRL is adopted to address challenges from the system complexity and uncertainty, which can work without the accurate system model and the distribution of uncertainties.
2) To guarantee the satisfaction of critical temperature constraints during DRL training, we adopt a safe-DRL framework. Specifically, we design a novel safe layer based on a linear program to achieve safety-imposing projection. This safe layer can project unsafe actions into safe ones to guide the agent's learning. As a result, it can effectively ensure constraint safety and protect the system from undesirable "trial-and-errors".
3) A self-adaption target method is proposed and designed as the reward function in the safe-DRL framework during the power recovery stage. The proposed method can effectively achieve smooth power recovery and avoid peak-power rebound that probably brings secondary impacts to power systems.

Besides, numerical studies verify the effectiveness of our proposed strategy, based on a real-world DCS. The analysis shows DCSs are qualified to provide operating reserve with mild impacts on buildings' indoor thermal comforts, subject to critical power constraints.

The rest is organized as follows. Section II introduces the physical architecture and control logic of DCS. Section III proposes the safe-DRL framework. Numerical studies are carried out in Section IV. Section V concludes this paper.

II. Modelling of the DCS

This section establishes the DCS model as the simulated environment to interact with the proposed DRL agent. Note that the only information received by the agent is the feedback from the established environment, but not details about the accurate DCS model.

A. DCS Framework

The schematic diagram of a DCS is shown in Fig. 2, in which blue lines represent the chilled water or cooling wind to supply cooling capacities for buildings; red lines are the returned warm water or warm wind. Its heat transmission process includes three isolated loops:

In the first water loop, chillers produce chilled water with a set temperature $T^{\mathrm{ch,s}}$, which is transported through pipelines to distributed buildings to supply cooling demands. The total mass flow m_t^{ch} is separated to different buildings by their independent two-port valves, which determine each building's own mass flow rate $m_{i,t}^{\mathrm{I}}$. After the heat exchange process, the chilled water in pipelines becomes warm with temperature $T_t^{\mathrm{ch,r}}$ and then is pumped back to chillers. The decoupler between the supply and return water balances pressure when the mass flow rate changes.

In the second water loop (i.e., water cycle in buildings), the water temperature $T_{i,t}^{\mathrm{II,s}}$ in buildings is cooled down by the chilled water in the first water loop through heat exchangers. Then the cool water transfers its thermal energy to the air in Air Handle Units (AHUs) to form cooling winds. The temperature of return water $T_{i,t}^{\mathrm{II,r}}$ reflects fluctuating cooling demands in buildings and further influences chillers' power consumption automatically.

In the air loop, AHUs blow cooling winds with the temperature $T_{i,t}^{\mathrm{w}}$ into each room, which can further influence the indoor temperature $T_{i,t}^{\mathrm{A}}$ and refresh the indoor air.

Note that the aforementioned three loops are mutually independent while interactional. Specifically, the total power consumption of a DCS majorly comes from chillers in the first water loop, whose operations are automatically and indirectly adjusted based on the buildings' cooling demands in the third loop. Therefore, it is significant to find the relationship between power consumption and these thermal dynamics.

B. Modelling of Key Components

1) Chillers: Chillers consume the most electricity in DCS. Their power consumption can be calculated based on the energy and mass balance, as follows:

$$P_t^{\mathrm{ch}} = \frac{Q_t^{\mathrm{ch}}}{\mathrm{COP}}, \quad \forall t, \tag{1}$$

This article has been accepted for publication in IEEE Transactions on Power Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2023.3237888

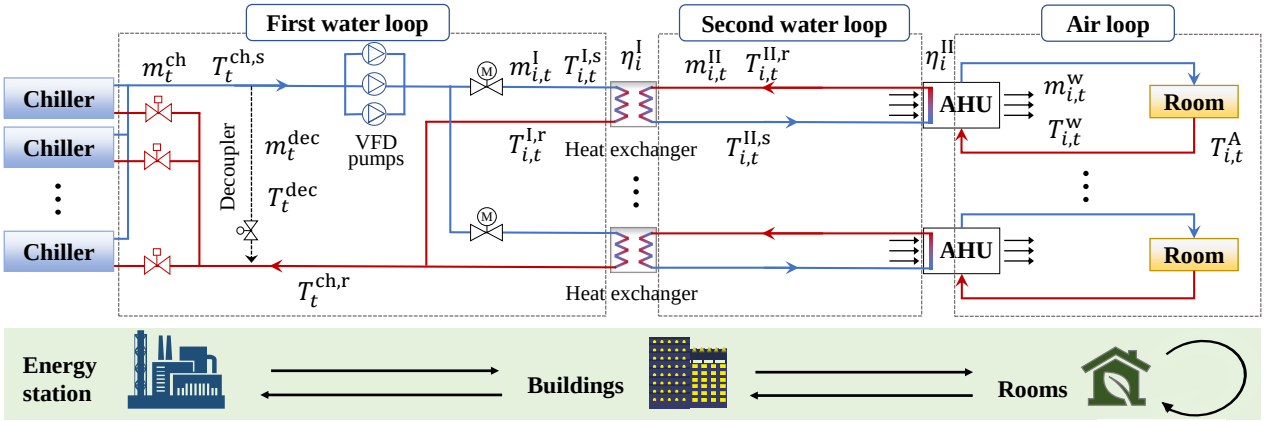


Fig. 2: Schematic diagram of a DCS.

where P_t^{ch} is chillers' electrical power at time t, in kW; Q_t^{ch} is the cooling power, in kW; COP denotes chiller's coefficient of performance. Generally, Q_t^{ch} is determined by chillers' varying return water temperature $T_t^{\text{ch,r}}$, in °C, and instantaneous mass flow rate m_t^{ch}, in kg/s, as follows:

$$Q_t^{\text{ch}} = m_t^{\text{ch}} c^{\text{w}} (T_t^{\text{ch,r}} - T^{\text{ch,s}}), \quad \forall t, \tag{2}$$

where c^{w} is the specific heat capacity of water, in kJ/(kg·°C). The set temperature of supply chilled water is represented by $T^{\text{ch,s}}$. Therefore, controlling the mass flow rate m_t^{ch} can influence the electrical power P_t^{ch} effectively. Further, we can rewrite $T_t^{\text{ch,r}}$ and m_t^{ch} according to the mass balance as:

$$m_t^{\text{ch}} = m_t^{\text{dec}} + \sum\nolimits_{i\in\mathcal{I}} m_{i,t}^{\text{I}}, \quad \forall t, \tag{3}$$

$$T_t^{\text{ch,r}} = \frac{m_t^{\text{dec}} c^{\text{w}} T_t^{\text{dec}} + \sum_{i\in\mathcal{I}} m_{i,t}^{\text{I}} c^{\text{w}} T_{i,t}^{\text{I,r}}}{m_t^{\text{ch}} c^{\text{w}}}, \quad \forall t, \tag{4}$$

where set $\mathcal{I}$ denotes the set of terminal buildings; $m_{i,t}^{\text{I}}$ and $T_{i,t}^{\text{I,r}}$ are each building's mass flow rate and return water temperature in first water loop, respectively; m_t^{dec} and T_t^{dec} are the mass flow rate and return water temperature of the decoupler, respectively. Eqs. (3)-(4) show the mass flow and energy balances between chillers and buildings.

2) Heat exchangers: Heat exchangers transfer the cooling supply from the first water loop to the second water loop. Considering the heat loss in pipelines, each building's actual supply water temperature can be calculated by:

$$T_{i,t}^{\text{I,s}} = T_t^{\text{out}} + \eta^{\text{pipe}} (T^{\text{ch,s}} - T_t^{\text{out}}), \quad \forall i \in \mathcal{I}, \forall t, \tag{5}$$

where η^{pipe} is the heat transfer coefficient of supply pipelines; T_t^{out} is the ambient temperature; $T_{i,t}^{\text{I,s}}$ is the supply chilled water temperature for building i. Further, the corresponding exchanging heat in building i, $Q_{i,t}^{\text{HE}}$ in kW, can be given by:

$$\begin{aligned} Q_{i,t}^{\text{HE}} &= m_{i,t}^{\text{II}} c^{\text{w}} (T_{i,t}^{\text{II,r}} - T_{i,t}^{\text{II,s}}) \\ &= \eta_i^{\text{I}} m_{i,t}^{\text{I}} c^{\text{w}} (T_{i,t}^{\text{I,r}} - T_{i,t}^{\text{I,s}}), \quad \forall i \in \mathcal{I}, \forall t, \end{aligned} \tag{6}$$

where η_i^{I} indicates the transfer efficiency of the first water loop to the second water loop; $m_{i,t}^{\text{I}}$ and $m_{i,t}^{\text{II}}$ are the mass flow rate of two sides, respectively. Similarly, $T_{i,t}^{\text{I,r}}, T_{i,t}^{\text{I,s}}$ and $T_{i,t}^{\text{II,r}}, T_{i,t}^{\text{II,s}}$ are the return and supply water temperature of each side, respectively. In addition, $Q_{i,t}^{\text{HE}}$ is determined by the performance of the heat exchanger, which can be calculated by [32]:

$$\frac{Q_{i,t}^{\text{HE}}}{k_i^{\text{HE}}} = \int_0^{F^{\text{HE}}} \Delta T_{i,t} \mathrm{d}F_i \approx F_i^{\text{HE}} \Delta T_{i,t}^{\text{mean}}, \quad \forall i \in \mathcal{I}, \forall t, \tag{7}$$

where k_i^{HE} is heat exchangers' transfer coefficient, in kW/(m^2·°C); F_i^{HE} is the surface area, in m^2; Symbol $\Delta T_{i,t}^{\text{mean}}$ is defined as the mean difference between the water's temperature of two sides, which is a function formulated as:

$$\Delta T_{i,t}^{\text{mean}} = \frac{(T_{i,t}^{\text{II,r}} - T_{i,t}^{\text{I,s}}) - (T_{i,t}^{\text{II,s}} - T_{i,t}^{\text{I,r}})}{\ln((T_{i,t}^{\text{II,r}} - T_{i,t}^{\text{I,s}})/(T_{i,t}^{\text{II,s}} - T_{i,t}^{\text{I,r}}))}, \quad \forall i \in \mathcal{I}, \forall t. \tag{8}$$

The above Eqs. (6)-(8) determine the dynamic exchanging heat in each building between the first and second water loops.

3) Buildings: AHU transfers the heat from the second water loop to the third air loop by blowing cooling wind, whose energy balance is given as:

$$m_{i,t}^{\text{w}} c^{\text{A}} (T_{i,t}^{\text{A}} - T_{i,t}^{\text{w}}) = \eta_i^{\text{II}} m_{i,t}^{\text{II}} c^{\text{w}} (T_{i,t}^{\text{II,r}} - T_{i,t}^{\text{II,s}}), \forall i \in \mathcal{I}, \forall t, \tag{9}$$

$$T_{i,t}^{\text{w}} = \frac{1}{2}(1-\alpha_i)(T_{i,t}^{\text{II,s}} + T_{i,t}^{\text{II,r}}) + \alpha_i T_t^{\text{out}}, \quad \forall i \in \mathcal{I}, \forall t, \tag{10}$$

where c^{A} and $m_{i,t}^{\text{w}}$ are air's specific heat capacity and wind's mass flow rate, respectively; η_i^{II} is the exchanging heat efficiency of second water loop to AHU; $T_{i,t}^{\text{A}} and T^{\text{out}}$ are the indoor and ambient temperature, respectively; $T_{i,t}^{\text{w}}$ represents the temperature of the cooling air out from AHU, mixing the outdoor fresh air with proportion α_i. Then the indoor thermal dynamic is described as:

$$c^{\text{A}} \rho^{\text{A}} V_i \frac{\mathrm{d}T_{i,t}^{\text{A}}}{\mathrm{d}t} = Q_{i,t}^{\text{loss}} - Q_{i,t}^{\text{DCS}}, \quad \forall i \in \mathcal{I}, \forall t, \tag{11}$$

where ρ^{A} is the density of the air, in kg/m^3; V_i is the space volume of the ith building, in m^3; $Q_{i,t}^{\text{loss}}$ is the ith building's heat loss because of its heat exchange with the ambient environment; $Q_{i,t}^{\text{DCS}}$ is the ith building's cooling gain from DCS, which are given as:

$$Q_{i,t}^{\text{DCS}} = m_{i,t}^{\text{w}} c^{\text{A}} (T_{i,t}^{\text{A}} - T_{i,t}^{\text{w}}), \quad \forall i \in \mathcal{I}, \forall t, \tag{12}$$

$$Q_{i,t}^{\text{loss}} = U_i^{\text{O-A}} A_i^{\text{S}} (T_t^{\text{out}} - T_{i,t}^{\text{A}}) + \zeta_{i,t}, \quad \forall i \in \mathcal{I}, \forall t, \tag{13}$$

where $U_i^{\text{O-A}}$ is the heat transfer coefficient, in kW/(m^2·°C); A_i^{S} is the surface area of the ith building, in m^2; $\zeta_{i,t}$ is the heat load from indoor sources (e.g., stochastic human behaviors and electric equipment), in kW.

The above models from Eq. (1) to Eq. (13) describe the whole thermal dynamics in a DCS. In summary, a DCS provides a cooling supply to multiple buildings through two water loops and one air loop by transmitting thermal energies.

Remark 1. *The chillers' cooling power is not only determined by the mass flow rate m_t^{ch} but also the uncertain return water temperature $T_t^{ch,r}$. The latter is further influenced by stochastic ambient temperature T^{out} and heat load ζ_i of buildings in Eq. (13). Besides, the accurate thermal model parameters in three loops are unknown and difficult to obtain in practice, which makes the conventional model-based control strategy infeasible for a DCS. To deal with these challenges, a model-free DRL method is proposed in the following Section III.*

III. Control of the DCS based on Safe Reinforcement Learning

This paper assumes that the DCS's promised reserve capacity is known in advance, e.g., offered by the DCS operator in the day-ahead or hour-ahead market.[1] This paper focuses on the controlling of the DCS to provide the promised reserve capacity. The DCS control objective during services is to provide high-quality performance and ensure all the buildings' indoor temperature comforts.

A. Formulation of the DCS Control Problem

In a practical DCS control problem, the next DCS operating state depends only upon the present state and the uncertain environment (e.g., the ambient temperature and building cooling demands), which satisfies the Markov property. Besides, the DCS state transition is independent of the index time t, so it satisfies the time-homogeneous property. Therefore, the DCS control problem is a typical sequential decision-making problem that can be described as an MDP [33].

In an MDP framework, a centralized smart controller, called *agent*, is designed to send each building signals to control its mass flow rate $m_{i,t}^{\text{I}}$. When a DCS provides operating reserve during the period $\mathcal{T} = [t_0, t_1]$, the DCS is regarded as an *environment*. Its real-time operation *state* $\boldsymbol{s}_t$ at time slot $t \in \mathcal{T}$ is observed by the agent. Then according to the information in $\boldsymbol{s}_t$, the agent makes one decision for DCS to execute *action* $\boldsymbol{a}_t$, which means there is a complete trajectory $\tau = \{\boldsymbol{s}_{t_0}, \boldsymbol{a}_{t_0+1}, \boldsymbol{s}_{t_0+1}, ..., \boldsymbol{a}_{t_1}, \boldsymbol{s}_{t_1}\}$ to describe the control process. The probability from the current state $\boldsymbol{s}_t$ to the next state $\boldsymbol{s}_{t+1}$ after taking action $\boldsymbol{a}_t$ is defined by a transition function $P(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t, \boldsymbol{a}_t)$, which is assumed unknown in model-free methods.

In the DCS control process, its power consumption and buildings' indoor temperature are the main considerations. The temperature deviation $\Delta T_{i,t} = T_{i,t}^{\text{A}} - T_{i,t}^{\text{set}}, \forall i \in \mathcal{I}, t \in \mathcal{T}$, is defined as the temperature comfort indicator, in which $T_{i,t}^{\text{set}}$ is the set temperature. Thus, the state and the action are defined by:

$$\boldsymbol{s}_t = \left[\Delta P_t, m_{i,t}^{\text{I}}, T_{i,t}^{\text{I,r}}, \Delta T_{i,t} | i \in \mathcal{I}\right]^{\intercal} \in \mathcal{S}, \quad \forall t \in \mathcal{T}, \quad (14)$$

$$\boldsymbol{a}_t = \left[\Delta m_{1,t}^{\text{I}}, \Delta m_{2,t}^{\text{I}}, ..., \Delta m_{|\mathcal{I}|,t}^{\text{I}}\right]^{\intercal} \in \mathcal{A}, \quad \forall t \in \mathcal{T}, \quad (15)$$

where ΔP_t equals to the gap between the actual power P_t^{ch} and required power cap P^{cap} of power systems. The scale of the state space $\mathcal{S}$ and action space $\mathcal{A}$ are $|\mathcal{S}| = 3|\mathcal{I}| + 1$ and $|\mathcal{A}| = |\mathcal{I}|$, respectively. The designed action space chooses the practical control variable to influence the DCS operating power, which is a complete space that can increase or decrease the mass flow to achieve optimal control. Here, actions can be executed directly through controlling valves in the system. Because the mass flow can be regulated by valves continuously, the action space is a continuous space and $\Delta m_{i,t}^{\text{I}}$ is a continuous variable. The positive (or negative) $\Delta m_{i,t}^{\text{I}}$ means to increase (or decrease) the mass flow rate, in which there are the upper bound $\overline{m}_i^{\text{I}}$ and lower bound $\underline{m}_i^{\text{I}}$ on the operating mass flow rate $m_{i,t}^{\text{I}}, \forall t$, in a real DCS. Therefore, the action value at each time step t satisfies the inequality $\underline{m}_i^{\text{I}} \leq \Delta m_{i,t}^{\text{I}} + m_{i,t-1}^{\text{I}} \leq \overline{m}_i^{\text{I}}$.

The designed state space captures the necessary system information about the control objective, control process, and environment uncertainties, which are strongly relevant to the decision variable. Specifically, the power deviation ΔP_t and building temperature deviation $\Delta T_{i,t}$ reflect whether the control results are good. The current mass flow $m_{i,t}^{\text{I}}$ is an important observation for the control process, because the action directly determines it. Besides, the return water temperature $T_{i,t}^{\text{I,r}}$ reflects buildings' cooling demands caused by the time-varying environment. Note that the ambient temperature is a weak-relevant variable in our problem, because the operating reserve is quite short (i.e., 15 minutes) so that the temperature variance is not significant to influence the agent's decision. Thus, it is not considered to prevent the unnecessary scale increase of the state space.

An arbitrary mapping from the state space to the action space $\pi : \mathcal{S} \rightarrow \mathcal{A}$ is called a *policy*. Essentially, the agent's task is to find an *optimal policy* that will be used as a guide for future online controlling. In order to evaluate a policy's performance, r_{t+1} is defined as *reward* for the action $\boldsymbol{a}_t$ in one step, which is formulated as:

$$r_{t+1} = -\theta^{\text{r}} \mathbb{E}_{i \in \mathcal{I}}[|\Delta T_{i,t+1}|] - \sigma_{i \in \mathcal{I}}^2[\Delta T_{i,t+1}], \ \forall t \in \mathcal{T}. \quad (16)$$

Eq. (16) includes two parts: the average and variance of all the buildings' temperature deviation at the next time $t+1$. Considering that the power requirement is a critical constraint for power systems imposed by the safe layer, we only consider temperature comfort as the control objective in the reward function. The former item $\mathbb{E}_{i \in \mathcal{I}}[|\Delta T_{i,t+1}|]$ is the average temperature of all the buildings' indoor temperature deviations from their corresponding set values. A smaller average value means fewer temperature influences on buildings. The later item $\sigma_{i \in \mathcal{I}}^2[\Delta T_{i,t+1}]$ is the variance of all the buildings' temperature deviations, where a smaller variance means less

[1]The capacity evaluation and strategic bidding for the DCS need to consider the energy cost, operating cost and market revenue together to maximize the economic benefit, which is out of the scope of this paper.

This article has been accepted for publication in IEEE Transactions on Power Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2023.3237888

difference of influences to different buildings. Parameter θ^{r} is the weight factor to determine the importance of the two parts. Note that in a real-time control problem, the economic benefit is not considered during the service. It is because the energy cost is decreased when DCS decreases its operating power to provide services, and the operating cost is not increased significantly due to the system's physical limitation.

Further, compared with the immediate reward r_t, the *return* G_t is defined as the accumulated reward in the future, which considers not only the immediate reward but also the expected influence on future rewards caused by the current action. The total discounted reward at time slot t is expressed as:

$$G_t = r_{t+1} + \gamma r_{t+2} + ... = \sum_{\tau=0}^{t_1-t} \gamma^{\tau} r_{t+\tau+1}, \ \forall t \in \mathcal{T}, \quad (17)$$

where $\gamma \in [0, 1]$ is a discount factor to represent the weight of the influence to future rewards [34]. For instance, when $\gamma = 1$, the agent considers the immediate and future rewards with the same importance. By contrary, when $\gamma = 0$, the agent only considers the current reward and $G_t = r_t$ Then, an *action-value function* $Q^{\pi}(\boldsymbol{s}_t, \boldsymbol{a}_t)$ is defined as the expected return from state $\boldsymbol{s}_t$, taking action $\boldsymbol{a}_t$ and following policy π:

$$Q^{\pi}(\boldsymbol{s}_t, \boldsymbol{a}_t) = \mathbb{E}_{\pi}[G_t | \boldsymbol{s}_t, \boldsymbol{a}_t], \quad \forall t \in \mathcal{T}, \quad (18)$$

where the *optimal action-value function* $Q^*(\boldsymbol{s}_t, \boldsymbol{a}_t)$ means the maximum action-value overall policies $\max_{\pi} Q^{\pi}(\boldsymbol{s}_t, \boldsymbol{a}_t)$. According to the theorem in MDP [22], optimal policy π^* is defined to satisfy $Q^{\pi^*}(\boldsymbol{s}_t, \boldsymbol{a}_t) = Q^*(\boldsymbol{s}_t, \boldsymbol{a}_t)$. Therefore, the agent's objective is to maximize the expected return J^{π}:

$$\max_{\pi} J^{\pi} = \mathbb{E}_{s_t \sim \mathcal{S}, a_t \sim \pi}[G_t] = \mathbb{E}_{s_t \sim \mathcal{S}}[Q^{\pi}(\boldsymbol{s}_t, \pi(\boldsymbol{s}_t))]. \quad (19)$$

However, different from the conventional policy optimization problem, there is a critical power constraint for a DCS during the power reduction stage and formulated as:

$$P_t^{\mathrm{ch}} \leq P^{\mathrm{cap}}, \quad \forall t \in \mathcal{T}, \quad (20)$$

where P^{cap} is the required power cap from the power system operator to constrain DCS's operating power. If it is violated, the DCS may be heavily penalized by the power system operator. Therefore, it is supposed to be satisfied at every time step. This critical constraint (20) turns the DCS control problem from a conventional MDP into an MDP with constraints.

B. Deterministic Deep Policy Gradient (DDPG) Algorithm

To solve the optimal policy π^* in Eq. (19), a safe-DDPG algorithm is proposed as shown in Fig. 3, which combines the actor-critic framework and deep Q-learning. Two neural networks are adopted to represent the action-value function Q and policy π, with parameters θ^Q and θ^{π}, respectively. The network to approximate Q value is called *critic network*, and another one that outputs actions is called *actor network*. In Fig. 3, the agent firstly interacts with the DCS environment to obtain transitions $(\boldsymbol{s}_t, \boldsymbol{a}_t, r_t, \boldsymbol{s}_{t+1})$, and collects all transitions into an *experience reply buffer* R. Secondly, the agent randomly samples a mini-batch of data from R to update two networks. Finally, the DCS receives an action that is produced by the actor-network $\pi(\boldsymbol{s}_t)$ and fine-tuned by the safe layer.

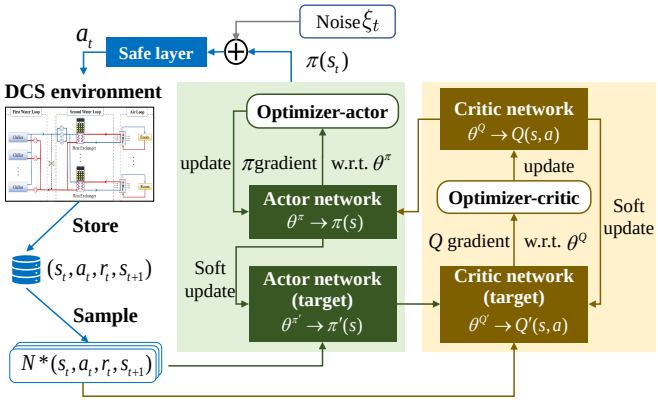


Fig. 3: Scheme of the safe-DDPG algorithm.

Using the experience reply buffer R, randomly sampled data keeps weak correlations with each other, which effectively avoids the over-fitting of two networks. In our work, we adopt the deterministic policy due to its high efficiency of sampling and computing compared with the stochastic policy [35]. For a deterministic policy, its gradient need not the integration of the action space like the stochastic policy, because the policy's action is deterministic [22]. According to the theorem proved by Silver [36], the policy gradient formulation is given as:

$$\begin{aligned} \bigtriangledown_{\theta^{\pi}} J^{\pi} &= \int_{\mathcal{S}} \rho^{\pi}(s) \bigtriangledown_{\theta^{\pi}} \pi(a|s) \bigtriangledown_{a} Q^{\pi}(s,a)|_{a=\pi(s)} \mathrm{d}s \\ &= \mathbb{E}_{s \sim \rho^{\pi}}[\bigtriangledown_{\theta^{\pi}} \pi(s) \bigtriangledown_{a} Q(s, \pi(s))], \end{aligned} \quad (21)$$

where the gradient of Q needs to be estimated through the critic network, and ρ^{π} is the state visitation distribution of policy π. Moreover, to give the unbiased estimation of the above gradient using the agent's sampled transitions, we can rewrite Eq. (21) based on the classical Monte-Carlo Simulation approach as follows:

$$\bigtriangledown_{\theta^{\pi}} J^{\pi} \approx \frac{1}{K} \sum_{k=1}^{K} \bigtriangledown_{\theta^{\pi}} \pi(\boldsymbol{s}_k) \bigtriangledown_{\boldsymbol{a}} Q(\boldsymbol{s}_k, \pi(\boldsymbol{s}_k)), \quad (22)$$

where k is the index of samples; K is the size of the sampled mini-batch data set $\mathcal{K}$. For the critic network θ^Q, the mean squared error (MSE) is used as the loss function:

$$L = \frac{1}{K} \sum_{k=1}^{K} [y_k - Q(\boldsymbol{s}_k, \boldsymbol{a}_k)]^2, \quad (23)$$

where y_k is the target value of $Q(\boldsymbol{s}_k, \boldsymbol{a}_k)$ and needs to be estimated. To stabilize the training process and guarantee convergence, the target y_k should not change frequently. According to the Bellman Expectation Equation of Eq. (18), two target networks (Q', π'), copies of ordinary networks (Q, π), are designed to calculate y_k as:

$$Q^{\pi}(\boldsymbol{s}_t, \boldsymbol{a}_t) = \mathbb{E}[r_t + \gamma Q^{\pi}(\boldsymbol{s}_{t+1}, \boldsymbol{a}_{t+1})], \quad \forall t \in \mathcal{T}, \quad (24)$$

$$y_k = r_k + Q'(\boldsymbol{s}_{k+1}, \pi'(\boldsymbol{s}_{k+1})), \quad \forall k \in \mathcal{K}. \quad (25)$$

To be more stabilized, the target networks are updated following the exponential moving average method:

$$\theta^{Q'} \leftarrow \tau \theta^{Q} + (1 - \tau) \theta^{Q'}, \quad (26)$$

$$\theta^{\pi'} \leftarrow \tau \theta^{\pi} + (1 - \tau) \theta^{\pi'}, \quad (27)$$

where τ is the smooth factor, $0 \leq \tau \ll 1$. Finally, to improve

This article has been accepted for publication in IEEE Transactions on Power Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2023.3237888

TABLE I: Safe-DDPG algorithm

01	Initialize the random process ξ, the experience reply buffer R and the actor, critic networks $Q(\boldsymbol{s},\boldsymbol{a}),\pi(\boldsymbol{s})$ with weights θ^Q,θ^π, respectively. Initialize corresponding two target networks Q',π' with weights $\theta^{Q'}\leftarrow\theta^Q$, $\theta^{\pi'}\leftarrow\theta^\pi$.
02	**For** episode = $1:1:M$ **do**
03	Receive initial observation state $\boldsymbol{s}_{t_0}$.
04	**For** $t=1:1:T$ **do**
05	Select DCSs' control action $\boldsymbol{a}_{t_0+t}=\pi(\boldsymbol{s}_{t_0+t})+\xi_{t_0+t}$.
06	Fine tune $\boldsymbol{a}_{t_0+t}$ by the safe layer.
07	Execute the action $\boldsymbol{a}_{t_0+t}$, then obtain the reward r_{t_0+t} and the next state $\boldsymbol{s}_{t_0+t+1}$.
08	Collect the transition $(\boldsymbol{s_{t_0+t}},\boldsymbol{a_{t_0+t}},\boldsymbol{r_{t_0+t}},\boldsymbol{s_{t_0+t+1}})$ to R, and randomly sample a mini-batch data from R.
09	Update the actor and critic networks by (25) and (23).
10	Update the two target networks by (26)-(27).
11	**Endfor**
12	**Endfor**

the efficiency of the exploration, an independent noise ξ_t is added to each action subject to the Gaussian distribution $\xi\sim N(0,\sigma^2)$. The proposed algorithm is summarized in Table I, where the safe layer shown as row 06 will be described in detail in the next subsection.

C. Constrained Policy by the Safe Layer

As shown in Fig. 3, the safe layer applies human oversight to the RL agents [31], which added two more steps before the action is executed. Specifically, at the first step, the safe layer intercepts unsafe actions and replaces them with safe ones. In the second step, the safe layer delivers a negative reward penalty for the agent choosing an unsafe action, which helps the agent learn to avoid unsafe actions in the future. Note that, by frequently replacing the agent's action with a different one, the safe layer essentially changes the state's transition function. It's conceivable that this added complexity makes the agent harder to learn a good policy. Hence, it is necessary to properly design the safe layer so that it not only ensures the safety of actions but also has a mild impact on the transition function to ensure good learning efficiency. The following part gives the design of our safe layer in detail.[2]

In each time step $t\in\mathcal{T}$, according to the output action $\boldsymbol{a}_t=\Delta\boldsymbol{m}_t^{\mathrm{I}}$, the next mass flow rate of buildings $\boldsymbol{m}_{t+1}^{\mathrm{I}}$ can be obtained as:

$$\boldsymbol{m}_{t+1}^{\mathrm{I}}=\boldsymbol{m}_t^{\mathrm{I}}+\Delta\boldsymbol{m}_t^{\mathrm{I}},\quad\forall t\in\mathcal{T}. \tag{28}$$

Thus, based on energy balance Eqs. (1)-(2), the power consumption at the next state is calculated as:

$$P_{t+1}^{\mathrm{ch}}=\sum\nolimits_{i\in\mathcal{I}}m_{i,t+1}^{\mathrm{I}}\Theta_t,\quad\forall t\in\mathcal{T}, \tag{29}$$

where $\Theta_t=\frac{1}{\mathrm{COP}}[c^{\mathrm{w}}(T_t^{\mathrm{ch,r}}-T^{\mathrm{ch,s}})]$ is the known parameter related with the return water temperature. Eq. (29) gives the explicit formulation of the power constraint (20) at each time step t. According to Eq. (29), we can judge whether the action is safe to satisfy the power constraint or not. If the power consumption satisfies the constraint $P_{t+1}^{\mathrm{ch}}\le P^{\mathrm{cap}}$, the action $\Delta\boldsymbol{m}_t^{\mathrm{I}}$ will be executed directly. Otherwise, it should be replaced by a safe one through the safe layer.

[2]In order to ensure the system operating safety, the equipment physical limits have been automatically guaranteed by the physical system through the saturation function. Therefore, the physical limits need not be considered as hard constraints in the safe layer.

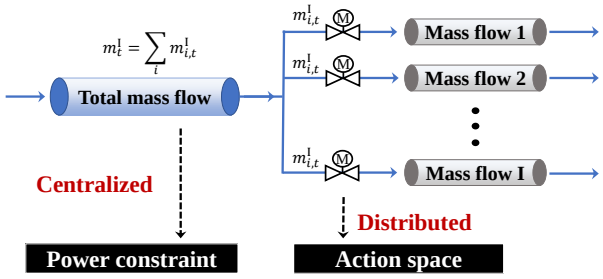


Fig. 4: Relationship between the action and constraint.

The relation between the power constraint and the action space is shown in Fig. 4, where the action $\Delta\boldsymbol{m}_{i,t}^{\mathrm{I}}$ is the adjustment of the mass flow to each building. When the constraint in Eq. (29) is not satisfied, the total mass flow needs to be decreased until satisfied. Then, the safe layer needs to revise all the buildings' actions (i.e., mass flows) to meet the required total mass flow. Note that a bad tuning rule will influence buildings' temperature comfort significantly, and then changes the agent's original action's reward. Therefore, by frequently replacing the agent's actions with the tuned ones, an improper safe layer may affect the learning convergence, because the agent's received rewards are not directly relevant to its original actions after tuning. Further, it is difficult to predict the temperature impact on each building caused by the adjustment of mass flows, because of the unknown dynamic thermal model and the uncertain cooling demands. Therefore, designing the tuning rule becomes not straightforward.

To address the above issue, we determine the tuning rule through two perspectives, which are the tuning direction and the tuning quantity. To adjust the unsafe action $\Delta\boldsymbol{m}_t^{\mathrm{I}}$ to a safe one $\Delta\tilde{\boldsymbol{m}}_t^{\mathrm{I}}$, the following linear mapping rule is proposed to determine the tuning direction:

$$\Delta\tilde{\boldsymbol{m}}_t^{\mathrm{I}}=\Delta\boldsymbol{m}_t^{\mathrm{I}}+\mu_t\Delta\boldsymbol{m}_t^{\mathrm{I}}+\upsilon_t\boldsymbol{m}_t^{\mathrm{I}},\quad\forall t\in\mathcal{T}, \tag{30}$$

where μ_t and υ_t are the correction coefficients, and $\mu_t,\upsilon_t\le 0$; $\Delta\tilde{\boldsymbol{m}}_t^{\mathrm{I}}$ is the updated action that will finally be executed in DCS. In Eq. (30), the relative relationship among buildings can be remained to ensure a similar temperature influence on buildings, so that the modification keeps the characteristic of the original action as much as possible. Coefficients μ_t and υ_t determine the tuning quantity of the original action. When μ_t and υ_t are close to 0, the last two correction terms in Eq. (30) will take a small function, i.e., the original agent's action $\Delta\boldsymbol{m}_t^{\mathrm{I}}$ will not be adjusted too much by the safe layer. By contrast, when μ_t and υ_t are negative and far from 0, the original agent's action $\Delta\boldsymbol{m}_t^{\mathrm{I}}$ will be adjusted significantly. Besides, the safe layer is not only a simple saturation function, but also needs to help the agent converge. If the decision from the agent $\Delta\boldsymbol{m}_t^{\mathrm{I}}$ is changed quite a lot by the safe layer, it may probably lead to failure in the agent's convergence. Therefore, the coefficients μ_t and υ_t are expected to be large and close to 0. On this basis, the two coefficients μ_t and υ_t can be optimized by the

This article has been accepted for publication in IEEE Transactions on Power Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2023.3237888

TABLE II: Safe layer method

01	Obtain the next mass flow rate $\boldsymbol{m}^{\mathrm{I}}_{t+1}$ and operating power $P^{\mathrm{ch}}_{t\ t+1}$ by (28), (29).
02	**If** $P^{\mathrm{ch}}_{t\ t+1} \le P^{\mathrm{cap}}$ **then**: execute $\Delta\boldsymbol{m}^{\mathrm{I}}_t$ directly;
03	**Else**
04	Solve the optimal coefficients μ_t and υ_t by (31)-(34);
05	Optimize the next mass flow rate $\Delta\boldsymbol{m}^{\mathrm{I}}_t$ using (30);
06	Execute the fine-tuned mass flow rate $\Delta\tilde{\boldsymbol{m}}^{\mathrm{I}}_t$.
06	**End**

following linear programming:

$$\max_{\mu_t,\upsilon_t}\ \mu_t+\upsilon_t, \tag{31}$$

$$\text{s.t.: } \sum\nolimits_{i\in\mathcal{I}}((\mu_t+1)\Delta m^{\mathrm{I}}_{i,t}+(\upsilon_t+1)m^{\mathrm{I}}_{i,t})\Theta_t \le P^{\mathrm{cap}},\ \forall t\in\mathcal{T}, \tag{32}$$

$$\underline{m}^{\mathrm{I}}_i \le (\mu_t+1)\Delta m^{\mathrm{I}}_{i,t}+(\upsilon_t+1)m^{\mathrm{I}}_{i,t} \le \overline{m}^{\mathrm{I}}_i,\ \forall i\in\mathcal{I},\forall t\in\mathcal{T}, \tag{33}$$

$$\mu_t,\upsilon_t,\le 0,\quad \forall t\in\mathcal{T}, \tag{34}$$

where the objective in Eq. (31) represents the minimum changes on the original agent's action $\Delta\boldsymbol{m}^{\mathrm{I}}_t$. The constraint in Eq. (32) is to satisfy the required power cap from power systems. Inequalities (33)-(34) define the domain of parameters $\mu_t,\upsilon_t\le 0$ and mass flow rate limitations $\underline{m}^{\mathrm{I}}_i \le m^{\mathrm{I}}_{i,t+1} \le \overline{m}^{\mathrm{I}}_i$. The calculation process of the safe layer is illustrated in Table II to achieve the fine-tuning of unsafe actions.

After the tuning, the safe layer will deliver a negative reward penalty to the agent for learning. More adjustment of the original action brings a higher penalty to the agent, which prevents the agent from relying heavily on the intervention of the safe layer to achieve safety control. The specific penalty is proportional to the adjustment coefficients, as follows:

$$r_{t+1} = -\theta^{\mathrm{r}}\mathbb{E}_{i\in\mathcal{I}}[|\Delta T_{i,t+1}|] - \sigma^2_{i\in\mathcal{I}}[\Delta T_{i,t+1}] - \theta^{\mathrm{safe}}|\mu_t+\upsilon_t|,\ \forall t\in\mathcal{T}. \tag{35}$$

Remark 2. *The mass flow rate is fine-tuned by a mapping rule as Eq. (30) to satisfy the power constraint. The tuning rule retains the original action's information as much as possible, which keeps all the buildings' relative relationship of mass flow rates. In this way, the safe layer can avoid causing significant influence on the convergence of the policy iteration.*

D. Self-adaptive Target Method

After providing operating reserve, DCS stops following power systems' regulation signals and enters the power recovery stage. Thus the power cap constraint in Eq. (20) is relaxed and the DCS tends to recover buildings' comfort temperature as soon as possible. However, a too rapid recovery of the temperature may cause an instantaneous increase in power consumption, called "power rebound". It may lead to a new power peak and cause stability problems for power systems. To avoid the "unsafe" power rebound, we further propose a self-adaption target method combined with the proposed safe-DDPG scheme to achieve a smooth recovery, as follows:

$$r_t = -\mathbb{E}_{i\sim\mathcal{I}}[|\Delta T_{i,t+1}-\varphi_{i,t+1}|],\quad \forall t\in[t_1,t_2], \tag{36}$$

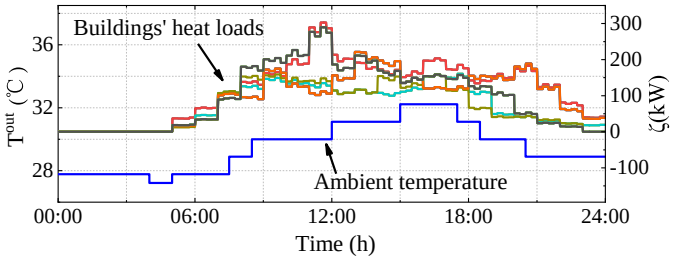


Fig. 5: The ambient temperature and buildings' heat loads.

where r_t is the reward of the indoor temperature in the recovery stage; $\varphi_{i,t}$ is the self-adaptive factor; t_1 is the end time of the power reduction stage and also the beginning time of the power recovery stage; t_2 is the required time for recovering the indoor temperature to the set value.

The reward r_t in Eq. (36) is different from the definition during the reduction stage in Eq. (16). Because Eq. (36) considers not only the buildings' set values, but also the self-adaptive factor $\varphi_{i,t}$ to design an expected temperature-decreasing trend. In this way, the sharp increase in the DCS's operating power can be alleviated. We propose the following configuration method for the self-adaptive factor[3]:

$$\varphi_{i,t} = \frac{\Delta T_{i,t_1}}{1+e^{\lambda[\frac{t-t_1}{t_2-t_1}-\frac{1}{2}]}},\quad \forall i\in\mathcal{I},\forall t\in[t_1,t_2], \tag{37}$$

where λ is determined according to the required recovery extent of the indoor temperature at time t_2. For example, when λ is set as 6, the recovery extent of the indoor temperature can reach 95% of $\Delta T_{i,t_1}$ at time t_2. Therefore, we can set the values of λ and t_2 to obtain the self-adaptive factor $\varphi_{i,t}$.

Moreover, in order to constrain the increased operating power during the recovery stage strictly, we also design a safe layer for the agent, similar to that during the reduction stage in Eq. (28)-(34). The difference is that the P^{cap} in Eq. (32) is replaced by the power consumption $P^{\mathrm{ch}}_{t_0}$ at time t_0 (i.e., the power consumption before the regulation), given by:

$$P^{\mathrm{ch}}_t \le P^{\mathrm{ch}}_{t_0},\quad \forall t\in[t_1,t_2], \tag{38}$$

where $\overline{P}^{\mathrm{ch}}$ is the upper limit of the operating power during the recovery stage.

Remark 3. *The proposed self-adaptive target method in Eq. (36) can regulate the DCS operating power to avoid the power rebound in the power recovery stage and minimize the buildings' comfort impacts.*

IV. Case Studies

A. Test System

The test system is modeled based on a realistic DCS in Hengqin, China, following its technical guidelines (the 4th Edition) [37]. The total installed cooling capacity in the energy station is 144 MW with COP = 5.5. The designed supply and return water temperature in two loops at time t=0 is $T^{\mathrm{ch,s}}=3$ °C, $T^{\mathrm{I,r}}_{i,0}=12$ °C, $T^{\mathrm{II,s}}_{i,0}=13$ °C, $T^{\mathrm{II,r}}_{i,0}=18$

[3]The configuration principle is to make the indoor temperature recover to 50% of $\Delta T_{i,t_1}$ when the time goes halfway, i.e., $t=t_1+\frac{1}{2}(t_2-t_1)$.

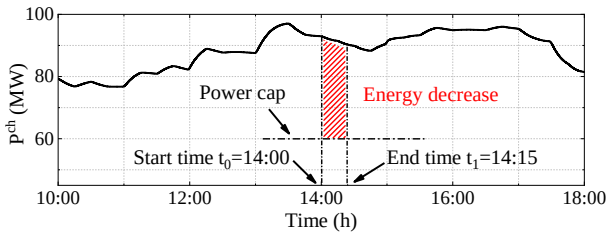


Fig. 6: The original power consumption of DCS.

°C, respectively. In addition, based on the national standard in China (JGJ 134-2010, GB 12021.3-2010, GB 31349-2014), the following parameters are designed as $k_i^{\text{HE}} = 4.5$ kW/(m^2·°C), $U^{\text{O-A}} = 0.0036$ kW/(m^2·°C), $c^{\text{w}} = 4.2$ kJ/(kg·°C), $c^{\text{A}} = 1.005$ kJ/(kg·°C) and $\rho^{\text{A}} = 1.205$ kg/m^3. The efficiency coefficients of the heat exchanging process between different loops are set as $\eta_i^{\text{I}} = 0.9$, $\eta_i^{\text{II}} = 0.9$, respectively. The heat transfer coefficient of supply water η^{pipe} is 0.95. The air mixing proportion is set as $\alpha_i = 0.1$.

The DCS in Hengqin provides cooling services for 12 buildings. The maximum value of the mass flow rate $\overline{m}_i^{\text{I}}$ ranges from 600 kg/s to 1,200 kg/s in different buildings, and the corresponding minimum value $\underline{m}_i^{\text{I}}$ is 3% of $\overline{m}_i^{\text{I}}$. Each building's floor area A_i^{S} and its set temperature $T_{i,t}^{\text{set}}$ are distributed in 100,000∼300,000 m^2, and 20∼23 °C, respectively. The maximum deviation of the required comfortable indoor temperature is ±1 °C. Moreover, the ambient temperature T_t^{out} and each building's heat load $\zeta_{i,t}$ adopt the realistic data in Hengqin, from June 1, 2020 to August 31, 2020 (one typical day's profiles are shown in Fig. 5).

The control objective of DCS is to provide operating reserve from 14:00 pm to 14:15 pm, as shown in Fig. 6.[4] The black curve is the original power consumption, and is regarded as the power baseline before regulation. The red shadow area is the required decrease in energy consumption, and the operating power should be lower than the power cap P^{cap}=60 MW during this period. In the recovery stage, the power before the regulation is set as the peak power of the baseline.

B. Benchmarks

To validate the superiority of the proposed safe-DDPG scheme, we implement another three control methods as our benchmarks: 1) the conventional proportional-integral (PI) controller [38]; 2) the MPC method [39]; 3) the conventional DRL controller [40]. Here, the superiority includes providing a higher-quality service and preventing the power rebound with minor impacts on buildings' temperature comforts.

In the PI method, the control signal is determined by the feedback of both the power cap violation and indoor temperature comfort. The buildings' mass flow regulation at each time step t can be expressed as:

$$\begin{aligned}\Delta m_t^{\text{ch}} =& \text{P}^{\text{ch}}(P_t^{\text{ch}} - P_{t-1}^{\text{ch}}) + \text{I}^{\text{ch}}(P_t^{\text{ch}} - P^{\text{cap}}), \qquad (39)\\ \Delta m_{i,t}^{\text{I}} =& \left[\text{P}_i(T_{i,t}^{\text{A}} - T_{i,t-1}^{\text{A}}) + \text{I}_i(T_{i,t}^{\text{A}} - T_{i,t}^{\text{set}})\right]\end{aligned}$$

[4]Note that these experimental settings are for illustrative purposes. In practice, the service duration, the operating reserve period and the power cap P^{cap} are determined by the system operator. The effectiveness of the proposed methodology is applicable to parameter settings in different scenarios.

TABLE III: Parameters for safe-DDPG and DDPG methods.

Symbols	Definitions	Values
τ	Target smooth factor	0.005
γ	Discount factor	0.9
$\|R\|$	Replay buffer capacity	10000
ξ	Exploration noise	0.3
M	Max episodes	2500
T	Max step	15
K	Mini batch size	200
δ_{θ^Q}	Learning rate of critic network Q	0.001
δ_{θ^π}	Learning rate of actor network π	0.0001
θ^{r}	Weight factor of temp deviations	0.01
θ^{p}	Weight factor of power violations	0.05
θ^{safe}	Weight factor of power violations	0.1

$$+ m_{i,t-1}^{\text{I}} \Delta m_t^{\text{ch}} / \sum\nolimits_{i\in\mathcal{I}} m_{i,t-1}^{\text{I}}, \qquad (40)$$

where P$^{\text{ch}}$, I$^{\text{ch}}$ are parameters of the PI controller in pipelines to follow power caps; P$_i$, I$_i$ are the parameters of PI controllers in buildings to follow set indoor temperatures. Eq. (40) means the regulation of the total mass flow is achieved by adjusting each building proportionally. During the power reduction stage, parameters are set as P$^{\text{ch}}$=0.2, I$^{\text{ch}}$=0.02. During the power recovery stage, parameters are set as P$^{\text{ch}}$=I$^{\text{ch}}$=0.

The MPC method requires a reliable predicted control model of the DCS thermal dynamic because it is a model-based method. Therefore, we assume that the MPC knows the accurate DCS dynamic model, as well as the distribution of the uncertainty. However, in practice, the model may be quite complex and challenging to obtain.

In the conventional DDPG method without a safe layer, the power constraint P^{cap} is considered as a penalty item in its reward function, which is formulated as:

$$\begin{aligned}r_{t+1} =& -\theta^{\text{r}}\mathbb{E}_{i\in\mathcal{I}}[|\Delta T_{i,t+1}|] - \sigma_{i\in\mathcal{I}}^2[\Delta T_{i,t+1}]\\ & - \theta^{\text{p}}|P_{t+1}^{\text{ch}} - P^{\text{cap}}|, \quad \forall t \in \mathcal{T}, \qquad (41)\end{aligned}$$

where θ^{p} is the weight factor of the penalty item.

C. Training Process of the Safe-DDPG Agent

The parameters of the proposed safe-DDPG are designed as Table III. The key hyper-parameters are designed based on the experience concluded in the existing literature, including the discount factor, learning rates, replay buffer capacity, etc. The actor and critic networks are composed by one input layer, two hidden layers and one output layer, respectively. The neurons number in each hidden layer is set as 128. The Rectified Linear Unit is used as the activation function. The parameters in DDPG (benchmark) adopt the same experimental settings as safe-DDPG. The simulation is implemented by the Windows system, using PyTorch in Python with an Intel core i7 CPU @3.0 GHz and 16GB memory.

The training process is shown in Fig. 7, and the number of training episodes is 2500. Fig. 7(a) presents the reward value for appraising the agent's decision in each episode. It can be seen that the rewards in safe-DDPG and DDPG have oscillations at first because of the unknown knowledge about the training environment (i.e., the DCS). With the increase of training episodes, the rewards converge to their respective

This article has been accepted for publication in IEEE Transactions on Power Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2023.3237888

TABLE IV: The training efficiency results for 2500 episodes.

Methods	Model	Training efficiency			Online solving time
		Sample efficiency	Convergence time	Convergence reward	
Safe-DDPG	model-free	∼500	2.9 min	-30	0.003 sec
DDPG	model-free	∼1000	4.3 min	-45	0.003 sec
MPC	model-based	–	–	–	2.76 sec

stable values, called the convergence reward. Then, both of the two agents obtain the optimal policy

The comparison of computational burden between three methods, the proposed safe-DDPG, the conventional DDPG, and the MPC method, is shown in Table IV. It can be seen that, although the MPC need not train its policy, it needs to resolve the optimization problem for every time step t. The average online solving time of the MPC is 2.76 seconds. For the two RL-based methods (i.e., safe-DDPG and DDPG), their policy networks need to be trained first, then their online solving time only costs 0.003 seconds on average, which is much shorter than that of the MPC. For the training process, sample efficiency is the estimated minimum number of samples to converge as illustrated in Fig. 7(a). It can be seen that the proposed safe-DDPG needs fewer sampled episodes to converge, which has higher sample efficiency and shorter convergence time. Besides, the convergence reward of safe-DDPG is larger (-30) than that of DDPG (-45), which means the proposed method can achieve the temperature objective better than DRL. It should be noted that both the reward curves of the safe-DRL and DRL use the first two items of their corresponding reward functions (i.e., Eq. (35) and Eq. (41)), which ensures the effectiveness of the comparison.

Fig. 7(b) shows the constraint violation during the agents' training processes, where $\Delta P = P^{\text{ch}} - P^{\text{cap}}$ is the power gap to the required power cap. It can be seen that the power constraint violation is conspicuous and even reaches over 40MW in DRL, which may harm the stable operation of the power system. However, the operating power can satisfy the power cap strictly in safe-DRL, which proves the effectiveness of the proposed safe layer. Besides, the operating power is quite close to the power cap, because the agent wants to make full use of the allowable power to decrease the indoor temperature deviations. Thus the well-trained agent can be applied to the online control of DCS for providing operating reserve.

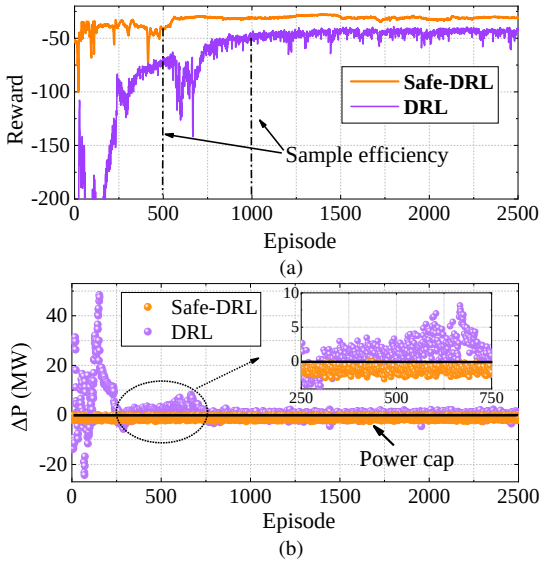


Fig. 7: The training process of the safe-DDPG agent. (a) The reward value; (b) The constraint value of operating power.

D. Online Control of DCS for Providing Operating Reserve

For a random case, it is assumed that the power system has the regulation demand at 14:00 pm, and sends the regulation signal to the agent to cut down the DCS's operating power to be lower than 60 MW in this dispatch period (15 min). The control results of the DCS for providing operating reserve are shown in Fig. 8, based on the four different controllers, i.e., PI in Fig. 8(a), the MPC method in Fig. 8(b), the conventional DDPG in Fig. 8(c) and the proposed safe-DDPG in Fig. 8(d).

It can be seen from Fig. 8(a) that DCS operating power is cut down and satisfies the required power after five min. Because the PI controller is designed based on the feedback, it cannot respond to the changing environment immediately which results in some time delay. In Fig. 8(b), Fig. 8(c) and Fig. 8(d), the three controllers can decrease the operating power more quickly compared with the PI controller, where the power reduction is achieved only within one min. Moreover, during the whole dispatch period, the operating power in Fig. 8(a) cannot be maintained below 60 MW and exceeds the required power cap at 14:11 due to the dynamic cooling demand in buildings (e.g., variational heat loads caused by people flows). By contrast, the operating power can be controlled under the power cap during the dispatch period in Fig. 8(b) and Fig. 8(d), which validates the effectiveness of the proposed safe-DDPG agent to satisfy power system's critical constraint strictly. Note that the MPC method can also solve a feasible solution to satisfy the power constraint strictly, because it has the accurate dynamic thermal model of DCS. In Fig. 8(c), the conventional DDPG method can also achieve the required power cap after training, however its training process in Fig. 7(b) can not satisfy the constraint.

After the power reduction stage, four controllers in Fig. 8 increase DCS's operating power to restore buildings' indoor temperatures. However, a new peak power of 114 MW and 104 MW appears in the recovery stage in Fig. 8(a) and Fig. 8(c), respectively. They are even much higher than the original daily maximum operating power (96 MW). This phenomenon may cause a secondary impact on the power system that has just returned to a stable state. In Fig. 8(b), the MPC also satisfies the power constraint well based on the known system model. In Fig. 8(d), the proposed safe layer limits the peak value during the recovery stage and guarantees the smooth recovery

This article has been accepted for publication in IEEE Transactions on Power Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2023.3237888

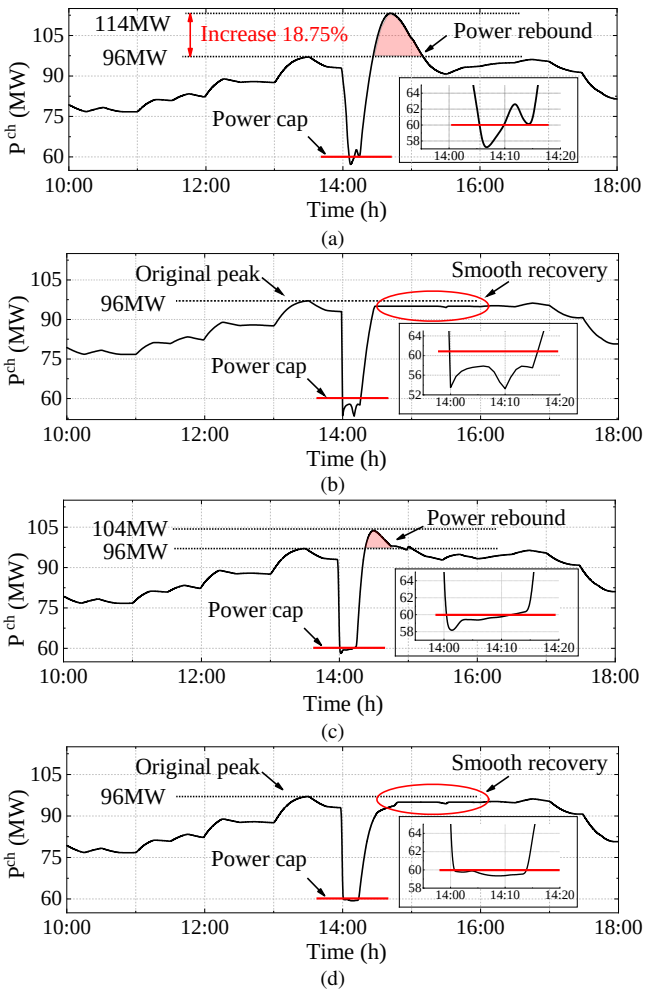


Fig. 8: The control power results of DCS based on (a) PI controller; (b) MPC method; (c) Conventional DDPG method; (d) Safe-DDPG method.

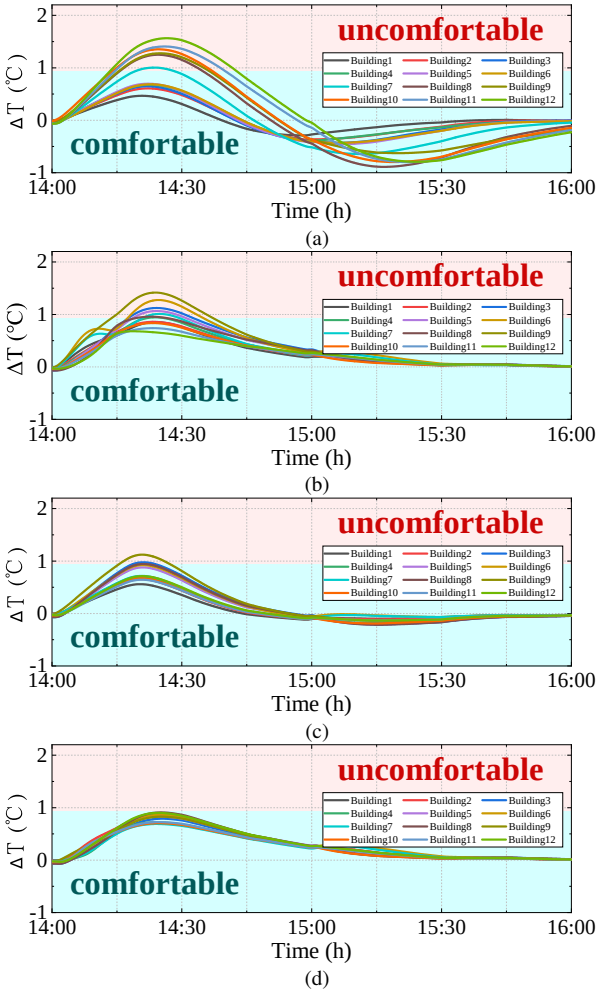


Fig. 9: The temperature deviation results of all the buildings based on (a) PI controller; (b) MPC method; (c) Conventional DDPG method; (d) Safe-DDPG method.

TABLE V: The statistical indicator of the temperature influence to buildings

Methods	Max deviation	Uncomfortable number	Average deviation
PI	1.57 °C	6	0.75 °C
MPC	1.41 °C	4	0.91 °C
DRL	1.18 °C	2	0.80 °C
Safe-DRL	0.93 °C	0	0.85 °C

of the operating power without a new peak power rebound.

Moreover, when DCS is controlled to provide the operating reserve, the building's indoor temperature will get influenced and deviates from its set value, as shown in Fig. 9. The blue area shows the comfortable temperature range in buildings, and ΔT denotes each building's temperature deviation. In the power reduction stage, all the buildings' indoor temperatures increase due to the reduction of cooling power supplies. In Fig. 9(a), more than half of the buildings' indoor temperatures deviate larger than 1 °C and enter the uncomfortable area. It means that some buildings get seriously impacted during the regulation process while others do not. In Fig. 9(b), the MPC does not achieve the temperature comfort in all the buildings, because the dynamic thermal model in DCS is nonlinear and makes it hard to converge to the optimum. In Fig. 9(c), the DDPG method can maintain the temperature comfort better than PI, while some buildings' indoor temperatures still exceed the comfortable range. By contrast, in Fig. 9(d), the temperature deviations in different buildings are close and maintained within 1 °C by using the safe-DDPG controller. As shown in Table V, it can be seen the max deviation of buildings based on safe-DDPG controller is the smallest and only equal to 0.93 °C. The uncomfortable number of buildings also decreases from 6 in the PI controller to 0 in safe-DDPG controller. The average temperature deviation of all the buildings distributes among 0.75 $\sim$ 0.85°C, which are similar and can be neglected for users' comfortable feelings. The above results validate the advantage of the proposed safe-DDPG method to regulate each building's mass flow rate dynamically and guarantee their temperature requirements.

V. Conclusion

This paper proposes a model-free safe-DDPG scheme for DCS control to provide the operating reserve. A safe layer is proposed to effectively guarantee the critical power constraint in the power reduction stage. A self-adaptive target method is adopted to tackle the power rebound in the power recovery

stage. Meanwhile, it minimizes the impacts on heterogeneous buildings' indoor temperatures to achieve the regulation within the required range $\pm$ 1°C. Numerical studies show that the DCS's operating power is always below the power cap during training, which ensures the "safety" of providing operating reserve. Besides, the DCS's operating power can recover smoothly and avoid undesirable peak power rebounds.

However, in our work, the promised capacity to the market is assumed known before the control, which is significant to assess the economic benefit of DCS [41]. The strategy to provide capacity offering considering energy cost, operating cost, and market revenue is beyond the scope of this paper, but will be our future work. Besides, the duration of operating reserve is assumed given in this work. If the duration changes, the influence brought by the operating reserve on buildings' temperature comfort will also change. Exploring the maximal buildings' regulation potential by utilizing their "thermal inertia" for different durations will be also our next work direction.

REFERENCES

[1] S. Impram, S. V. Nese, and B. Oral, "Challenges of renewable energy penetration on power system flexibility: A survey," *Energy Strategy Rev.*, vol. 31, p. 100539, Sep. 2020.

[2] H. Nosair and F. Bouffard, "Reconstructing operating reserve: Flexibility for sustainable power systems," *IEEE Trans. Sustain. Energy*, vol. 6, no. 4, pp. 1624–1637, Oct. 2015.

[3] P. Siano, "Demand response and smart grids—a survey," *Renew. Sust. Energ. Rev.*, vol. 30, pp. 461–478, Feb. 2014.

[4] M. Cai, M. Pipattanasomporn, and S. Rahman, "Day-ahead building-level load forecasts using deep learning vs. traditional time-series techniques," *Appl. Energy*, vol. 236, pp. 1078–1088, Feb. 2019.

[5] S. Werner, "International review of district heating and cooling," *Energy*, vol. 137, pp. 617–631, Oct. 2017.

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

[7] PJMINT.L.L.C., "PJM manual 11: Energy&ancillary services market operations," Revision: 115, pp. 83-98, Jun. 01, 2021. [Online]. Available: https://www.pjm.com/-/media/documents/manuals/m11.ashx .

[8] C. Cheng, B. Yang, and F. Wang, "Research on the application of large scale air conditioning load in demand response," *Power Demand Side Management*, vol. 19, no. 3, pp. 57–59, Apr. 2017.

[9] G. Chen, B. Yan, H. Zhang, D. Zhang, and Y. Song, "Time-efficient strategic power dispatch for district cooling systems considering the spatial-temporal evolution of cooling load uncertainties," to appear in CSEE J. Power Energy Syst., 2021. DOI: 10.17775/CSEE-JPES.2020.06800.

[10] C.-C. Lo, S.-H. Tsai, and B.-S. Lin, "Ice storage air-conditioning system simulation with dynamic electricity pricing: A demand response study," *Energies*, vol. 9, no. 2, p. 113, 2016.

[11] R. Tang, S. Wang, K. Shan, and H. Cheung, "Optimal control strategy of central air-conditioning systems of buildings at morning start period for enhanced energy efficiency and peak demand limiting," *Energy*, vol. 151, pp. 771–781, 2018.

[12] M. D. Knudsen and S. Petersen, "Model predictive control for demand response of domestic hot water preparation in ultra-low temperature district heating systems," *Energy Build.*, vol. 146, pp. 55–64, Jul. 2017.

[13] D. M. Alghool, T. Y. Elmekkawy, M. Haouari, and A. Elomri, "Optimization of design and operation of solar assisted district cooling systems," *Energy Convers. Manag.: X*, vol. 6, p. 100028, Apr. 2020.

[14] A. Stoppato, A. Benato, N. Destro, and A. Mirandola, "A model for the optimal design and management of a cogeneration system with energy storage," *Energy Build.*, vol. 124, pp. 241–247, Jul. 2016.

[15] Z. Zhang, D. Zhang, and R. C. Qiu, "Deep reinforcement learning for power system applications: An overview," *CSEE Journal of Power and Energy Systems*, vol. 6, no. 1, pp. 213–225, 2019.

[16] K. Mason and S. Grijalva, "A review of reinforcement learning for autonomous building energy management," *Computers & Electrical Engineering*, vol. 78, pp. 300–312, 2019.

[17] S. Qiu, Z. Li, Z. Li, J. Li, S. Long, and X. Li, "Model-free control method based on reinforcement learning for building cooling water systems: Validation by measured data-based simulation," *Energy and Buildings*, vol. 218, p. 110055, 2020.

[18] S. Liu and G. P. Henze, "Evaluation of reinforcement learning for optimal control of building active and passive thermal storage inventory," 2007.

[19] Y. Du, H. Zandi, O. Kotevska, K. Kurte, J. Munk, K. Amasyali, E. Mckee, and F. Li, "Intelligent multi-zone residential hvac control strategy based on deep reinforcement learning," *Appl. Energy*, vol. 281, p. 116117, Jan. 2021.

[20] X. Xu, Y. Jia, Y. Xu, Z. Xu, and C. S. Lai, "A multi-agent reinforcement learning-based data-driven method for home energy management," *IEEE Trans. Smart Grid*, vol. 11, no. 4, pp. 3201–3211, Jul. 2020.

[21] L. Yu, Y. Sun, Z. Xu, C. Shen, D. Yue, T. Jiang, and X. Guan, "Multi-agent deep reinforcement learning for HVAC control in commercial buildings," *IEEE Trans. Smart Grid*, vol. 12, no. 1, pp. 407–419, Jan. 2021.

[22] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*, Sep. 2016.

[23] R. Dobbe, P. Hidalgo-Gonzalez, S. Karagiannopoulos, R. Henriquez-Auba, G. Hug, D. S. Callaway, and C. J. Tomlin, "Learning to control in power systems: Design and analysis guidelines for concrete safety problems," *Electr. Power Syst. Res.*, vol. 189, p. 106615, Dec. 2020.

[24] F. Yongjie, "Reflections on frequency stability control technology based on the blackout event of 9 august 2019 in uk," *Automation of Electric Power Systems*, vol. 43, no. 24, pp. 1–5, 2019.

[25] K. P. Schneider, E. Sortomme, S. S. Venkata, M. T. Miller, and L. Ponder, "Evaluating the magnitude and duration of cold load pick-up on residential distribution using multi-state load models," *IEEE Transactions on Power Systems*, vol. 31, no. 5, pp. 3765–3774, 2016.

[26] J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained policy optimization," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 06–11 Aug 2017, pp. 22–31. [Online]. Available: https://proceedings.mlr.press/v70/achiam17a.html

[27] C. Tessler, D. J. Mankowitz, and S. Mannor, "Reward constrained policy optimization," *CoRR*, vol. abs/1805.11074, 2018. [Online]. Available: http://arxiv.org/abs/1805.11074

[28] T. Yang, J. Rosca, K. Narasimhan, and P. J. Ramadge, "Projection-based constrained policy optimization," *CoRR*, vol. abs/2010.03152, 2020. [Online]. Available: https://arxiv.org/abs/2010.03152

[29] T. M. Moldovan and P. Abbeel, "Safe exploration in markov decision processes," *CoRR*, vol. abs/1205.4810, 2012. [Online]. Available: http://arxiv.org/abs/1205.4810

[30] G. Dalal, K. Dvijotham, M. Vecerík, T. Hester, C. Paduraru, and Y. Tassa, "Safe exploration in continuous action spaces," *CoRR*, vol. abs/1801.08757, 2018. [Online]. Available: http://arxiv.org/abs/1801.08757

[31] W. Saunders, G. Sastry, A. Stuhlmüller, and O. Evans, "Trial without error: Towards safe reinforcement learning via human intervention," *CoRR*, vol. abs/1707.05173, 2017. [Online]. Available: http://arxiv.org/abs/1707.05173

[32] N. Marshall, "Heat exchange design handbook," *Int. J. Heat Fluid Flow*, vol. 4, no. 2, p. 77, 1983.

[33] O. Alagoz, H. Hsu, A. J. Schaefer, and M. S. Roberts, "Markov decision processes: a tool for sequential decision making under uncertainty," *Med. Decis. Making*, vol. 30, no. 4, pp. 474–483, Dec. 2010.

[34] ——, "Markov decision processes: a tool for sequential decision making under uncertainty," *Med. Decis. Making*, vol. 30, no. 4, pp. 474–483, 2010.

[35] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep reinforcement learning: A brief survey," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 26–38, 2017.

[36] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *International conference on machine learning*. PMLR, 2014, pp. 387–395.

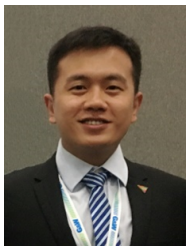
[37] New District Hengqin: District Cooling and Heating System Technical Guidelines (the 4th Edition), Zhuhai, China, Mar. 2016.

[38] Y. Chen, H. Yan, Y. Luo, and H. Yang, "A proportional–integral (PI) law based variable speed technology for temperature control in indirect evaporative cooling system," *Appl. Energy*, vol. 251, p. 113390, Oct. 2019.

[39] "Theory and applications of hvac control systems – a review of model predictive control (mpc)," *Building and Environment*, vol. 72, pp. 343–355, 2014.
[40] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.
[41] Z. Shi, C. Wang, X. Lei, X. Ye, W. Yuan, and Z. Cao, "Research on the technical economy and market mechanism of electric heat storage system participating in auxiliary service," in *2018 2nd IEEE Conference on Energy Internet and Energy System Integration (EI2)*, Oct. 2018, pp. 1–6.

Yonghua Song (F'08) received the B.E. and Ph.D. degrees from the Chengdu University of Science and Technology, Chengdu, China, and the China Electric Power Research Institute, Beijing, China, in 1984 and 1989, respectively, all in electrical engineering. He was awarded DSc by Brunel University in 2002, Honorary DEng by University of Bath in 2014 and Honorary DSc by University of Edinburgh in 2019. From 1989 to 1991, he was a Post-Doctoral Fellow at Tsinghua University, Beijing. He then held various positions at Bristol University, Bristol, U.K.; Bath University, Bath, U.K.; and John Moores University, Liverpool, U.K., from 1991 to 1996. In 1997, he was a Professor of Power Systems at Brunel University, where he was a Pro-Vice Chancellor for Graduate Studies since 2004. In 2007, he took up a Pro-Vice Chancellorship and Professorship of Electrical Engineering at the University of Liverpool, Liverpool. In 2009, he joined Tsinghua University as a Professor of Electrical Engineering and an Assistant President and the Deputy Director of the Laboratory of Low-Carbon Energy. During 2012 to 2017, he worked as the Executive Vice President of Zhejiang University, as well as Founding Dean of the International Campus and Professor of Electrical Engineering and Higher Education of the University. Since 2018, he became Rector of the University of Macau and the director of the State Key Laboratory of Internet of Things for Smart City. His current research interests include smart grid, electricity economics, and operation and control of power systems. Prof. Song was elected as the Vice-President of Chinese Society for Electrical Engineering (CSEE) and appointed as the Chairman of the International Affairs Committee of the CSEE in 2009. In 2004, he was elected as a Fellow of the Royal Academy of Engineering, U.K. In 2019, he was elected as a Foreign Member of the Academia Europaea.

Peipei Yu (S'21) received the M.S and B.S. degree in mathematics from Zhejiang University, Zhejiang, China, in 2019 and 2016, respectively. She is currently working toward the Ph.D. degree at University of Macau, Macau, China. Her research interests include Internet of Things for smart energy, demand response and reinforcement learning control.



Hongxun Hui (S'17–M'20) received both the Ph.D. and B. Eng degrees in electrical engineering from Zhejiang University in 2020 and 2015, respectively. He is currently a Post-doctoral Fellow with the State Key Laboratory of Internet of Things for Smart City, University of Macau. His research interests include modelling and optimal control of demand side resources in smart grid, the electricity market considering demand response, and the uncertainty analysis brought by flexible loads and renewable energies.

Hongcai Zhang (S'14–M'18) received the B.S. and Ph.D. degree in electrical engineering from Tsinghua University, Beijing, China, in 2013 and 2018, respectively. He is currently an Assistant Professor with the State Key Laboratory of Internet of Things for Smart City and Department of Electrical and Computer Engineering, University of Macau, Macao, China. In 2018-2019, he was a postdoctoral scholar with the Energy, Controls, and Applications Lab at University of California, Berkeley, where he also worked as a visiting student researcher in 2016. His current research interests include Internet of Things for smart energy, optimal operation and optimization of power and transportation systems, and grid integration of distributed energy resources.

Ge Chen (S'20) received the B.S. degree from Huazhong University of Science and Technology, Wuhan, China, in 2015 and the M.S. degree from Xi'an Jiaotong University, both in thermodynamic engineering. He is currently working toward the Ph.D. degree at University of Macau, Macau, China. His research interests include Internet of Things for smart energy, optimal operation and data-driven optimization under uncertainty.