entropy

MDPI

Article

Energy Dispatch for CCHP System in Summer Based on Deep Reinforcement Learning

Wenzhong Gao and Yifan Lin *

Merchant Marine College, Shanghai Maritime University, 1550 Haigang Avenue, Pudong District, Shanghai 201306, China
* Correspondence: 202030110029@stu.shmtu.edu.cn; Tel.: +86-183-2110-0872

Abstract: Combined cooling, heating, and power (CCHP) system is an effective solution to solve energy and environmental problems. However, due to the demand-side load uncertainty, load-prediction error, environmental change, and demand charge, the energy dispatch optimization of the CCHP system is definitely a tough challenge. In view of this, this paper proposes a dispatch method based on the deep reinforcement learning (DRL) algorithm, DoubleDQN, to generate an optimal dispatch strategy for the CCHP system in the summer. By integrating DRL, this method does not require any prediction information, and can adapt to the load uncertainty. The simulation result shows that compared with strategies based on benchmark policies and DQN, the proposed dispatch strategy not only well preserves the thermal comfort, but also reduces the total intra-month cost by 0.13~31.32%, of which the demand charge is reduced by 2.19~46.57%. In addition, this method is proven to have the potential to be applied in the real world by testing under extended scenarios.

Keywords: deep reinforcement learning; DoubleDQN; CCHP system; energy dispatch; demand charge; uncertainties

Citation: Gao, W.; Lin, Y. Energy Dispatch for CCHP System in Summer Based on Deep Reinforcement Learning. *Entropy* **2023**, *25*, 544. https://doi.org/10.3390/e25030544

Academic Editors: Andrea Prati, Luis Javier García Villalba and Vincent A. Cicirello

Received: 16 February 2023
Revised: 12 March 2023
Accepted: 17 March 2023
Published: 21 March 2023

1. Introduction

The utilization of fossil fuels has caused environmental problems globally, such as air pollution and climate warming [1]. At the same time, energy consumption rises instead of falling with the rapid depletion of energy resources, of which about 40% is used for the production of cool, heat, and electricity [2]. Compared with the traditional energy supply system, the combined cooling, heating, and power (CCHP) system can promote the coordinated operation of the above three kinds of energy, which provides an effective way to solve environmental and energy problems with high energy efficiency. As a consequence, CCHP systems have been widely used in residential and office buildings, and hospitals in the past decade [3–5].

Although high energy efficiency has been verified in engineering applications, existing CCHP systems usually suffer from high operating costs, especially in summer. For example, Shanghai, a city in southern China, has a subtropical monsoon climate with outdoor temperatures up to 39 °C in summer, so a large amount of electricity is required to provide cooling to balance the demand-side user's high cooling demand, which leads to the surge of CCHP system's energy purchasing cost. Therefore, it is very necessary to optimize existing systems to achieve optimal economic operation in summer while meeting the demand-side energy load.

In the related literature, there are many studies focusing on optimizing the dispatch strategy of the CCHP system. Ref. [6] established a dispatch model of the CCHP system with the optimization goal of energy cost and carbon emission, and solved it by CPLEX solver. Ref. [7] established a two-stage dispatch model for the CCHP system based on quantity adjustment and proposed an iterative solution algorithm. Ref. [8] proposed an improved bee colony algorithm to solve the multi-objective dispatching model, so as to

balance economic benefits and environmental friendliness. Ref. [9] built a load prediction model using neural networks, and proposed a feed-forward active operation optimization method considering load prediction for the CCHP system. Refs. [10–12] improved the traditional electric load following strategy and thermal load following strategy to improve the matching degree of energy output and demand. Ref. [13] used a genetic algorithm to optimize the dispatch strategy of the CCHP system to minimize the energy cost under different operating circumstances. In addition, because of human activities and outdoor weather conditions, there are dynamic uncertainties in the demand-side load, which brings a severe challenge to the energy dispatch of the CCHP system. To solve this problem, some studies tried to optimize the CCHP system dispatch strategy using model predictive control (MPC) [14–16] and robust optimization [17–19] separately, and the dispatch result showed that high-quality optimization depended on the accuracy of prediction.

Although the above studies have laid the solid foundation for the optimal energy dispatch of the CCHP system, the impact of demand charge on the operating cost is not considered. According to surveys, the so-called demand charge is already common among power industries in countries including China, the U.S, and Sweden, and is generally charged based on the electricity customers' monthly peak power demand to the grid, that is, the peak electric power purchase (in kW) regardless of timing, rather than the cumulatively purchased electricity (in kWh) [20,21]. Additionally, some studies suggest that introducing demand charge into the utility rate has two potential benefits: (1) incentivizing smarter demand-side management; (2) guaranteeing the stability of the grid, and avoiding power accidents that may endanger public safety, such as large-scale blackouts [22,23]. The presence of demand charge further increases the difficulty of CCHP system energy dispatching. Therefore, in view of the above two factors, this paper will achieve the optimal economic operation of the CCHP system in summer under the rate structure including time-of-use electricity price and demand charge.

On the other hand, the optimization methods applied in all the above study works can be categorized as model-based methods. Although the model-based method well reflects the thermodynamic performance of the CCHP system and the internal mechanism of demand-side load variation, it relies on the expert experience of modeling or the prediction information of future uncertainty, which is difficult to adapt to dynamic changes of the actual environment. Once the operating status and structure of the CCHP system change with time, the model-based method needs to remake the dispatch strategy, which increases the computational burden and greatly reduces the decision-making efficiency. Besides, the model-based method often suffers from a low intelligence level and a long solution time, which cannot meet the requirements for real-time operation.

Therefore, in view of the above limitations, this paper introduces deep reinforcement learning (DRL) into the CCHP system energy dispatching problem. DRL is an emerging technique that has received extensive attention in recent years. With its strong perception and decision-making ability, DRL not only ensures real-time requirements, but also provides a model-free optimization approach to generate adaptive strategies without prior knowledge of the environment, avoiding drawbacks of the model-based method. In this sense, the DRL-based method has more advantages.

Studies have demonstrated that DRL can be used to solve complex high-dimensional control and optimization problems such as in robots [24], games [25], and traffic controls [26]. In the field of energy dispatching, DRL-based methods have also attracted broad attention. Ref. [27] proposed a microgrid energy management method based on a deep Q network to achieve higher economic benefits under stochastic conditions. Ref. [28] achieved the optimal control of the heating, ventilation, and air conditioning (HVAC) system using deep deterministic policy gradient (DDPG) to reduce energy costs and thermal discomfort. After improving the traditional DDPG algorithm, ref. [29] proposed a dynamic dispatch method for an integrated energy system based on improved DDPG, and verified the superiority of this method. Ref. [30] used DRL to optimize the electricity dispatch of an all-electric ferry, so as to achieve the dual-objective optimization of energy cost and load expected

loss. Ref. [31] developed a microgrid energy management method using proximal policy optimization (PPO) to deal with the uncertainties of renewable energy. Ref. [32] developed a CHP system dispatch method using PPO, efficiently handling the wind-turbine failure without rewriting constraints. Although these study works have made significant contributions, they only considered a simple optimization objective of minimizing costs due to grid exchange and power generation, ignoring the impact of the demand charge mechanism and the control of peak exchanging power, which may result in an extra ultra-high cost. In addition, there is not only no benchmark policy designed to further verify the superiority and generalization of the proposed method, but also no comprehensive discussion of its potential application in real scenarios from multiple aspects.

Therefore, motivated by the above problems, this paper aims to develop a DRL-based energy dispatch method for the CCHP system to achieve optimal economic operation in summer. The innovation and main contributions are summarized as follows:

- We focus on the energy dispatch for the CCHP system in summer (EDCS) and formulate the EDSC problem as a Markov decision process (MDP), in which the load uncertainty, energy cost, demand charge, and energy balance are considered.
- DRL algorithm DoubleDQN is used to solve the formulated MDP and make dispatch strategies for the CCHP system. In contrast to previous study works, the proposed method directly makes decisions based on the current state, getting rid of the dependence on the accuracy of prediction information and model description.
- By comparing with the DQN-based method and benchmark policies, the advantages of the proposed method in computational efficiency, total intra-month cost saving, and peak power purchase control are verified.
- From the aspects of dealing with unseen physical environments, load fluctuation, and sudden unit failure, the potential of application in real scenarios is discussed.

The rest of the paper is organized as follows. The EDCS problem is mathematically formulated in Section 2; the proposed DRL-based method is introduced in Section 3; the case study is given in Section 4; finally, the conclusion is drawn in Section 5.

2. EDCS Problem Formulation

EDCS problem aims to efficiently manage the energy output of the CCHP system, to achieve optimal economic operation in summer. Therefore, this section first introduces a typical CCHP system, then establishes mathematical models of its key units, and finally designs the objective function and constraints of the EDCS problem.

2.1. System Description

2.1.1. Structure of CCHP System

Figure 1 shows a typical CCHP system, which consists of an internal combustion engine (ICE), absorption chiller (AC), electric chiller (EC), cooling tower (CT), gas boiler (GB), storage tank (ST) and auxiliary equipment (AE). The high-grade heat generated by the burning of natural gas is used to drive ICE to produce electricity, and the low-grade waste heat is used to produce cool through AC. If ICE fails to meet the electricity load, it will be supplemented by the grid. The cooling load is mainly satisfied by EC and AC, and the low-grade heat generated during the cooling process will be released through CT. The heating load is mainly satisfied by GB. Insufficient heating and cooling energy are supplied by ST, while the operation of this system is assisted by AE.

Additionally, the following assumptions are made for the typical CCHP system:

(1) Demand-side user is directly connected to the grid, and thus the electricity load only consists of ECs, AEs, and CT.

(2) ICEs and ECs run at rated power.

(3) ICEs and ACs run in a one-to-one matching mode, meaning the number of running ICEs is equal to the number of running ACs at each time step.

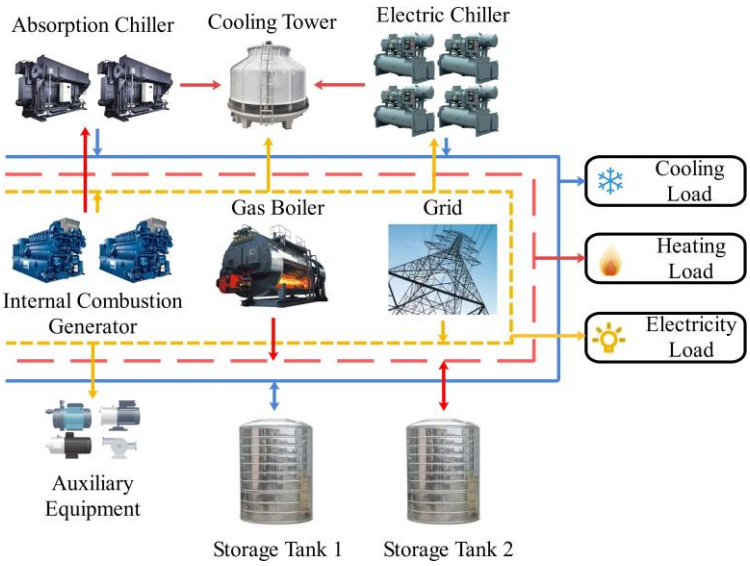

Figure 1. Structure and energy flow of the typical CCHP system.

2.1.2. Mathematical Model of Key Unit

Mathematical modeling is carried out for key units of the CCHP system under summer conditions, including ICE, AC, EC, ST, and AE:

(1) ICE

The ICE electric power output $P^{ICE}(t)$ (kW) at time step t is defined according to the following equation:

$$P^{ICE}(t) = \begin{cases} P^{ICE}_{rated}, \ if \text{ ICE } is\ running \\ 0, otherwise \end{cases} \quad (1)$$

where, P^{ICE}_{rated} is the rated electric power generation of ICE (kW). Then, the ICE gas consumption $V^{ICE}(t)$ (m^3/h) and the low-grade waste heat Q^{waste} (kW) can be calculated as follows:

$$V^{ICE}(t) = \frac{P^{ICE}(t)}{\eta_{ICE} \cdot \varepsilon_{LHV}} \quad (2)$$

$$Q^{waste}(t) = \left(\frac{1-\eta_{ICE}}{\eta_{ICE}}\right) \cdot P^{ICE}(t) \quad (3)$$

where, η_{ICE} is the electric efficiency of ICE and ε_{LHV} is the low calorific value of natural gas (kWh/m^3).

(2) AC

The AC cooling output $Q^{AC}(t)$ (kW) at time step t is jointly decided by the absorbed waste heat $Q^{waste}(t)$ and the coefficient of performance COP_{AC}, that is:

$$Q^{AC}(t) = Q^{waste}(t) \cdot COP_{AC} \quad (4)$$

(3) EC

The EC cooling output $Q^{EC}(t)$ (kW) at time step t is defined according to the following equation:

$$Q^{EC}(t) = \begin{cases} Q^{EC}_{rated}, \ if \text{ EC } is\ running \\ 0, otherwise \end{cases} \quad (5)$$

where, Q^{EC}_{rated} is the rated cooling generation of EC (kW). Then, the EC electric power consumption $P^{EC}_e(t)$ (kW) can be calculated as the following equation:

$$P^{EC}_e(t) = Q^{EC}(t)/COP_{EC} \quad (6)$$

where, COP_{EC} is the coefficient of performance of EC.

(4) ST

The energy storage relationship of ST between adjacent time steps is defined as the following equation:

$$E^{ST}(t) = E^{ST}(t-1) - Q^{ST}(t)\cdot\Delta t \tag{7}$$

where, Δt (h) is the time step interval, $E^{ST}(t)$ (kWh) and $Q^{ST}(t)$ (kW) are the remaining energy storage and the storing ($Q^{ST}(t) < 0$) or releasing ($Q^{ST}(t) \geq 0$) power of ST at time step t, respectively.

(5) CT and AE

Since the running of CT and AE are closely related to the operating status of other energy supply units, the electricity consumption of CT and AE is allocated to ECs, ACs, ICEs, and ST for the convenience of calculation, where ICE and AC are allocated as a whole because of the matching mode, that is:

$$P_e^{CT}(t) + P_e^{AE}(t) = P_{e,a}^{EC}(t) + P_{e,a}^{ICE\&AC}(t) + P_{e,a}^{ST}(t) \tag{8}$$

$$P_{e,a}^{EC}(t) = \begin{cases} \alpha_{EC}\cdot\left(n^{EC}(t)\right)^2 + \beta_{EC}\cdot n^{EC}(t) + \gamma_{EC}, if\ n^{EC}(t) > 0 \\ 0, otherwise \end{cases} \tag{9}$$

$$P_{e,a}^{ICE\&AC}(t) = \begin{cases} \alpha_{ICE}\cdot\left(n^{ICE}(t)\right)^2 + \beta_{ICE}\cdot n^{ICE}(t) + \gamma_{ICE},\ if\ n^{ICE}(t) > 0 \\ 0,\ otherwise \end{cases} \tag{10}$$

$$P_{e,a}^{ST}(t) = \begin{cases} \alpha_{ST}\cdot\left(Q^{ST}(t)\right)^2 + \beta_{ST}\cdot\left(Q^{ST}(t)\right) + \gamma_{ST}, \quad if\ \ Q^{ST}(t) > 0 \\ 0,\ otherwise \end{cases} \tag{11}$$

where, $P_e^{CT}(t)$ and $P_e^{AE}(t)$ are the electric power consumption (kW) of CT and AE, respectively; $P_{e,a}^{EC}(t)$, $P_{e,a}^{ICE\&AC}(t)$ and $P_{e,a}^{ST}(t)$ are the allocated electric power consumption (kW) of ECs, ACs, ICEs and ST, respectively; $n^{EC}(t)$ and $n^{ICE}(t)$ are the number of running ECs and running ICEs, respectively; α, β, γ are the electric power consumption coefficients.

2.2. Objective Function

Specifically, the objective of EDCS is to minimize the total intra-month cost C_{total}^{CCHP} (RMB), which is composed of the total energy cost C_{total}^{ECT} (RMB) and the total demand charge C_{total}^{DC} (RMB), by optimally dispatching each unit of the CCHP system. So, the objective function can be defined as follows:

$$min\ C_{total}^{CCHP} = min\left\{C_{total}^{ECT} + C_{total}^{DC}\right\} \tag{12}$$

C_{total}^{ECT} can be calculated by:

$$C_{total}^{ECT} = \sum_{t=1}^{T} c_t^{ECT} = \sum_{t=1}^{T}\left(\mu_{gas}\cdot\Delta t\cdot\sum_{i=1}^{n^{ICE}(t)} V_i^{ICE}(t) + \left(\begin{cases}\mu_{grid}(t)\cdot P^{grid}(t)\cdot\Delta t,\ if\ \ P_{grid}(t) > 0 \\ \mu_{sell}\cdot P^{grid}(t)\cdot\Delta t,\ otherwise\end{cases}\right)\right) \tag{13}$$

where, T (h) is the total time steps in a dispatch period, c_t^{ECT} (RMB) is the energy costs at time step t, $V_i^{ICE}(t)$ (m^3/h) is the gas consumption of ith running ICE; μ_{gas} (RMB/m^3), $\mu_{grid}(t)$ (RMB/kWh) and μ_{sell} (RMB/kWh) are the natural gas price, and the purchasing and selling electricity price, respectively.

C_{total}^{DC} is obtained according to [33], and can be calculated by:

$$C_{total}^{DC} = \mu_{demand}\cdot max_{1\leq t\leq T}\left\{\max\left(P^{grid}(t), 0\right)\right\} \tag{14}$$

where, $max_{1\le t\le T}\left\{\max\left(P^{grid}(t),0\right)\right\}$ (kW) is the peak electric power purchase in a dispatch period, μ_{demand} (RMB/kW) is the unit price of the demand charge. Further, to allocate C^{DC}_{total} to each time step [34], Equation (14) can be mathematically transformed into:

$$C^{DCS}_{total} = \sum_{t=1}^{T} c^{DCS}_t = \mu_{demand} \cdot \sum_{t=1}^{T}\left(\frac{t}{T-1}\cdot P^{peak}_t - \frac{t-1}{T-1}\cdot P^{peak}_{t-1}\right) \quad (15)$$

$$P^{peak}_t = max\left\{P^{peak}_{t-1}, max\left(0, P^{grid}(t)\right)\right\} \quad (16)$$

where, c^{DCS}_t (RMB) is the demand charge at time step t; P^{peak}_t is the peak electric power purchase (kW) in the last t time steps and defines $P^{peak}_0 = 0$.

Therefore, the EDCS problem objective function can be eventually expressed as:

$$min\ C^{CCHP}_{total} = min\left\{\sum_{t=1}^{T}(c^{EC}_t + c^{DCS}_t)\right\} \quad (17)$$

2.3. Constraints

2.3.1. Energy balance Constraints

The energy balance constraints of the CCHP system under summer conditions can be listed as follows.

Cooling balance:

$$\sum_{i=1}^{n^{EC}(t)} Q^{EC}_i(t) + \sum_{i=1}^{n^{ICE}(t)} Q^{AC}_i(t) + Q^{ST}(t) = Q_d(t) \quad (18)$$

where, $Q^{EC}_i(t)$ (kW) and $Q^{AC}_i(t)$ (kW) are the cooling output of ith running EC and ith running AC at time step t, respectively; $Q_d(t)$ is the cooling load (kW).

Electricity balance:

$$\sum_{i=1}^{n^{ICE}(t)} P^{ICE}_i(t) + P^{grid}(t) = P_d(t) = \sum_{i=1}^{n^{EC}(t)} P^{EC}_{e,i}(t) + P^{CT}_e(t) + P^{AE}_e(t) \quad (19)$$

where, $P^{ICE}_i(t)$ (kW) and $P^{EC}_{e,i}(t)$ (kW) are the electric power output and consumption of ith running ICE and ith running EC, respectively; $P^{grid}(t)$ (kW) is the exchanging of electric power with the grid, $P^{grid}(t) > 0$, if power is purchased, otherwise $P^{grid}(t) \le 0$; $P_d(t)$ is the electricity load (kW).

2.3.2. Operational Constraints

Besides energy balance constraints expressed in Equations (18) and (19), there are some operational constraints for energy supply units of the CCHP system.

ECs, ICEs, and ACs are constrained by the quantity. Hence,

$$0 \le n^{EC}(t) \le n^{EC}_{max} \quad (20)$$

$$0 \le n^{ICE}(t) \le n^{ICE}_{max} \quad (21)$$

$$0 \le n^{AC}(t) \le n^{AC}_{max} \quad (22)$$

where n^{EC}_{max}, n^{ICE}_{max} and n^{AC}_{max} are the maximum number of ECs, ICEs, and ACs, respectively.

ST is constrained by the capacity and the storing/releasing power. Hence,

$$0 \le E^{ST}(t) \le E^{ST}_{rated} \quad (23)$$

$$\left|Q^{ST}(t)\right| \le Q^{ST}_{max} \quad (24)$$

where, Q^{ST}_{max} is the maximum storing/releasing the power of ST; E^{ST}_{rated} is the rated capacity of ST and defines $E^{ST}(0) = 0$.

3. DRL-Based EDSC Method

Since the EDCS objective function designed in Section 2.2 is to determine the output of energy supply units at each time step so as to minimize the total intra-month cost, the EDCS problem is essentially a sequential decision-making problem. In this section, we first convert the EDCS problem into a Markov decision process (MDP), in which the energy balance constraints and operational constraints of the CCHP system are considered. And the DRL algorithm is adopted to find the optimal strategy for this MDP.

3.1. Converting of EDCS Problem into MDP

An MDP is usually defined as a tuple $\langle S, A, p, r \rangle$, where S is the state space, A is the action space, p is the state transition probability, $r: S \times A \to r$ is the reward function. The agent observes the current environment state $s \in S$ and chooses an action $a \in A(s)$, where $A(s)$ represents the set of all admissible actions at state s [35].

In this paper, the CCHP system is the environment where the agent is located, and the interaction between the agent and the environment is shown in Figure 2: at time step t, the agent observes the environment state s_t, and generates an action a_t based on the policy π (policy is a mapping from state s to action a, that is, $a = \pi(s)$) to make dispatch strategy so as to determine the output of energy supply units.

Figure 2. Interaction between the agent and the environment.

Next, we convert the EDCS problem into an MDP, and the fundamental elements of which are defined as follows.

(1) State

At time step t, the environment state information for the agent includes physical time, remaining ST storage, peak electric power purchase obtained so far, purchasing electricity price, and cooling load. Among them, the cooling load is a state variable with uncertainty, which can't be controlled by the agent. Thus, the state s_t (5-dimension) is described as:

$$s_t = \left[t, E^{ST}(t), P_{t-1}^{peak}, \mu_{grid}(t), Q_d(t) \right] \tag{25}$$

(2) Action

The aim of the EDCS problem is to decide the electric power output ($\sum_{i=1}^{n^{ICE}(t)} P_i^{ICE}(t) + P^{grid}(t)$) and the cooling output ($\sum_{i=1}^{n^{EC}(t)} Q_i^{EC}(t) + \sum_{i=1}^{n^{AC}(t)} Q_i^{AC}(t) + Q^{ST}(t)$) of CCHP system. In fact, after $n^{ICE}(t)$ is decided, $\sum_{i=1}^{n^{ICE}(t)} P_i^{ICE}(t)$ and $\sum_{i=1}^{n^{AC}(t)} Q_i^{AC}(t)$ can be determined by Equations (1) and (4), respectively; after $n^{EC}(t)$ is decided, $\sum_{i=1}^{n^{EC}(t)} Q_i^{EC}(t)$ can be determined by Equation (5). Further, $Q^{ST}(t)$ and $P^{grid}(t)$ can be calculated by Equations (18) and (19), respectively. In other words, the output of other energy supply units can be obtained

immediately after $n^{EC}(t)$ and $n^{ICE}(t)$ are jointly decided. Therefore, the agent's action a_t at time step t can be represented by $n^{EC}(t)$ and $n^{ICE}(t)$, that is:

$$a_t = \left[n^{EC}(t), n^{ICE}(t)\right] \tag{26}$$

where, $a_t \in A$, and A is the set of all admissible actions that satisfy the EDCS constraints. According to Equations (20) and (21), $[0,\ n^{EC}_{max}]$ and $[0, n^{ICE}_{max}]$ are the range of $n^{EC}(t)$ and $n^{ICE}(t)$, respectively.

(3) Reward function

According to the EDCS objective function and constraints, the goal of the agent is to minimize the total cost which is composed of the energy cost and demand charge, while balancing the supply and demand of energy. In order to achieve this goal, the reward r_t received by the agent consists of the above three parts, and can be defined as:

$$r_t = -\lambda_1 \cdot c_{EC}(t) + \left[-\lambda_2 \cdot c_{DC}(t) - \theta \cdot \max\left(0,\ P_t^{peak} - \varepsilon_{peak}\right)\right] - \lambda_3 \cdot \Delta Q_d(t) \tag{27}$$

where, λ_1, λ_2 and λ_3 are the weighting factors, $\Delta Q_d(t)$ is the cooling error (kW) (that is, the difference between supply and demand of cooling power); $\theta \cdot \max\left(0,\ P_t^{peak} - \varepsilon_{peak}\right)$ is the extra penalty obtained when the peak power purchase P_t^{peak} is larger than the threshold ε_{peak}. The setting of r_t transfers the total intra-month cost minimization problem to the reward maximization form of the MDP.

From the viewpoint of MDP, the quality of chosen action a at the given environment state s can be evaluated by the state-action value function $Q_\pi(s, a)$:

$$Q_\pi(s, a) = E_\pi\left[\sum_{k=0}^{T} \tau^k \cdot r_{t+k} | s_t = s, a_t = a\right] \tag{28}$$

where, $\tau \epsilon [0, 1]$ is the discount factor used to balance the future reward and current received reward [36], $E_\pi[\cdot]$ is the reward expectation under the policy π. The aim of the agent is to find an optimal policy π^*, so as to maximize the function $Q_\pi(s, a)$, that is, $\pi^* = argmax_{a \in A} Q_\pi(s, a)$.

On the other hand, the above MDP doesn't define the state transition probability p. With a full knowledge of p, the model-based method can solve $E_\pi\left[\sum_{k=0}^{T} \tau^k \cdot r_{t+k} | s_t = s, a_t = a\right]$ (that is, the Q_π value of action a) through fully observing the environment [37]. However, due to the influence of dynamic uncertainty factors such as human activity, environmental weather, and unit failure, the establishment of p model becomes extremely difficult, which makes the model-based method, not a suitable solution.

Therefore, in this paper, the model-free DRL-based method is used to solve the EDCS problem under the MDP framework. By interacting with the environment, the DRL-based method can incrementally improve its decision strategy without any information on state transition probability p.

3.2. DRL Solution

3.2.1. A Brief Review of DRL

Reinforcement learning (RL) is a paradigm of machine learning. With the assistance of the Q table, the RL algorithm can iteratively update the state-action value function based on the reward function defined in MDP [38]:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \psi \cdot [r(s_t, a_t) + \tau \cdot max_{a \in A} Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)] \tag{29}$$

where, $\psi \epsilon [0, 1]$ is the learning factor.

DRL is the combination of deep neural network (DNN) and RL [39], the essence of which is to approximate $Q(s, a)$ through the nonlinear function. When encountering problems with large state space, DRL utilizes DNN as the regression tool, which solves the

potential dimension-explosion problem of the RL algorithm caused by the establishment of a huge Q table [40].

Additionally, DRL can be generally divided into the value-based DRL algorithm for discrete action space and the policy-based DRL algorithm for continuous action space [37,41]. Since the action space defined in Equation (25) is discrete, the value-based DRL algorithm is adopted to generate a dispatch strategy for the CCHP system.

3.2.2. Basic Principles of Value-Based DRL Algorithm

A general DNN structure for the value-based DRL algorithm is shown in Figure 3. Specifically, DNN is used to evaluate the Q value of each potential action corresponding to the state s, and the agent will select the action with the highest Q value.

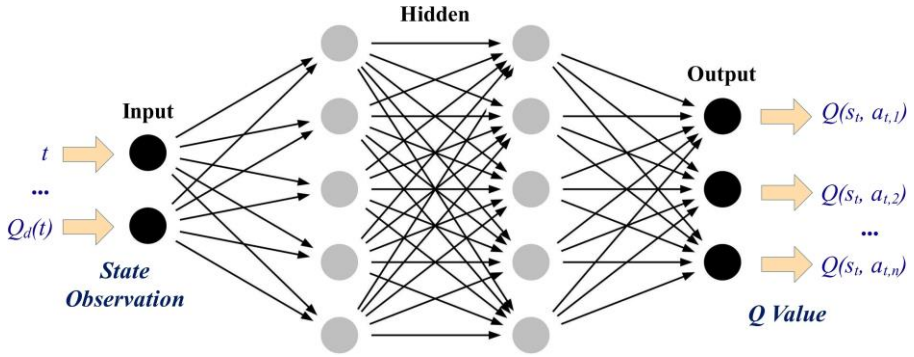

Figure 3. General DNN structure for the value-based DRL algorithm.

A deep Q network (DQN) is a representative value-based DRL algorithm [42], which has two DNNs named Q network and the target network. The training objective of DQN is to minimize the loss function $L(\omega)$:

$$\begin{cases} L(\omega) = E[y_t - \mathrm{Q}(s_t, a_t; \omega))]^2 \\ y_t = r_t + \tau \cdot max_a Q(s_{t+1}, a; \omega') \end{cases} \tag{30}$$

where, y_t is the target Q value, $y_t - \mathrm{Q}(s_t, a_t; \omega)$ is the time-difference error; ω and ω' are weight parameters of the Q network and target network, respectively.

However, as shown in Equation (30), Q values used for action evaluation and selection in DQN are both outputs by the target network, which tends to cause overvaluation. In order to solve this problem, a greedy-policy-based double deep Q network (DoubleDQN) algorithm is proposed to decouple the evaluation and selection [43]. DoubleDQN evaluates the Q value using the Q network and selects the action to take using the target network. The target Q value is then:

$$y_t = r_t + \tau \cdot \mathrm{Q}\left(s_{t+1}, argmax_a \mathrm{Q}(s_{t+1}, a; \omega); \omega'\right) \tag{31}$$

and the gradient $\nabla_\omega L(\omega)$ provides the direction for DNN parameters updating, that is:

$$\begin{cases} \nabla_\omega L(\omega) = E[2 \cdot (y_t - Q(s_t, a_t; \omega)) \cdot \nabla_\omega Q(s_t, a_t; \omega)] \\ \omega \leftarrow \omega - \psi \cdot \nabla_\omega L(\omega) \end{cases} \tag{32}$$

where, weights ω is updated every step and copied to weights ω' every fixed number of steps. Additionally, a mechanism called experience replay is integrated into DoubleDQN [44]. In this mechanism, the agent stores the experience $e_t = (s_t, a_t, r_t, s_{t+1})$ at each time step, and randomly extracts a batch of experience samples for the off-line training.

3.2.3. Realizing EDCS with DoubleDQN

The framework of the DoubleDQN-based EDCS method is shown in Figure 4. For the DoubleDQN network, the input is a 5-dimensional vector $s_t = \left[t, E^{ST}(t), P_{t-1}^{peak}, \mu_{grid}(t), Q_d(t)\right]$, and the outputs are the Q values of all potential actions.

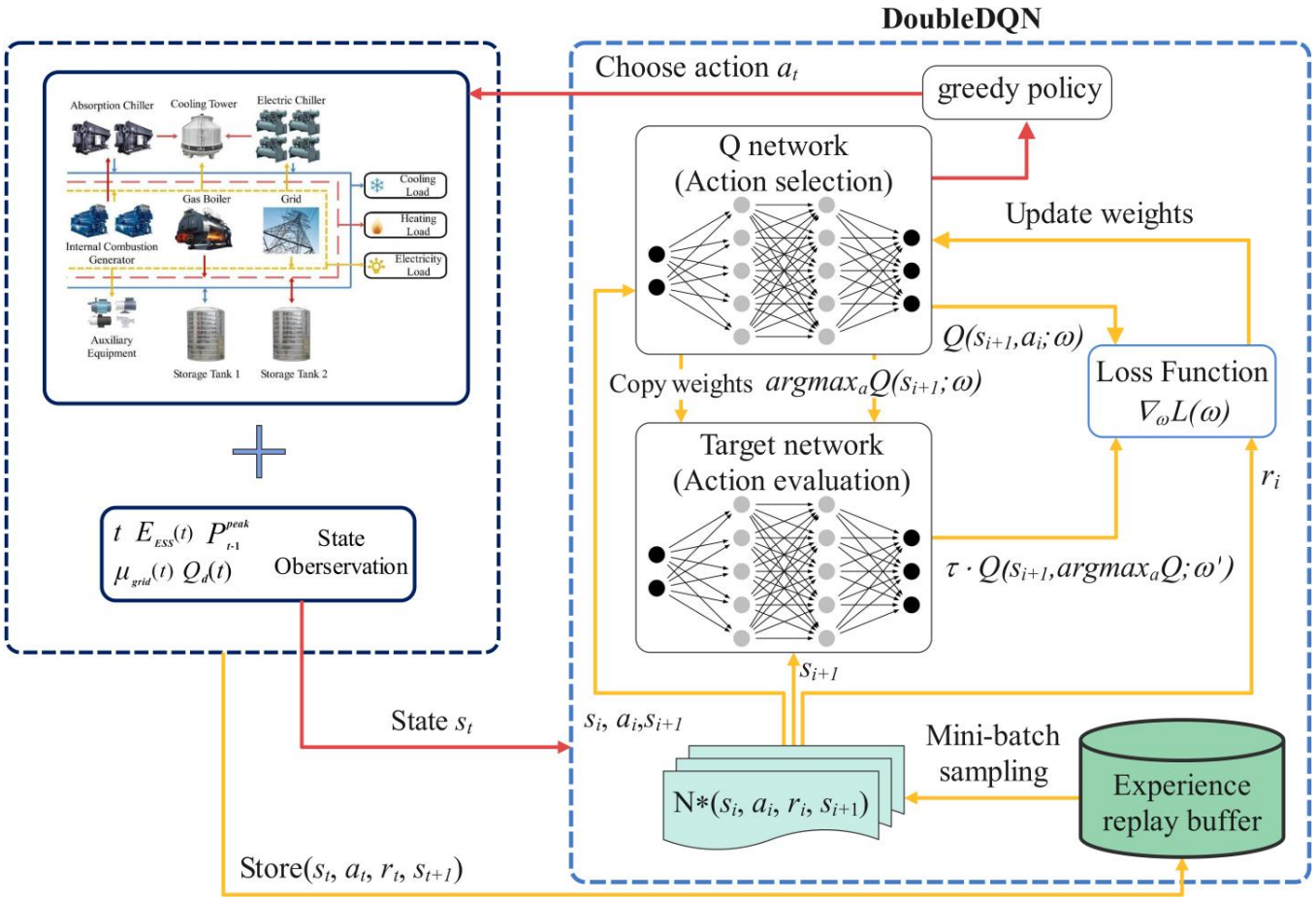

Figure 4. Framework of the proposed DoubleDQN-based EDCS method.

We use historical data of the CCHP system as environment states to train the DoubleDQN algorithm offline. Its input includes physical time, ST storage, peak electric power purchase, electricity price, and cooling load. After the offline training process shown in Algorithm 1, the parameters of DoubleDQN will be fixed and used for the online decision-making of the CCHP system.

The decision-making procedure of the proposed DoubleDQN-based EDCS method can be found in Algorithm 2. When the dispatch begins, the weights ω of the Q network trained by Algorithm 1 are loaded. At each time step t, the agent selects an action a_t based on the current CCHP state s_t. Next, the action is executed by energy supply units, and the CCHP environment transits to the next state s_{t+1}. Meanwhile, the agent receives the reward r_t and observes s_{t+1} as the current state. This procedure repeats until the end of the dispatch period. From the procedure, it can be found that the proposed method requires no prediction information, realizing a direct mapping from the real-time state observation to the CCHP system energy dispatching.

Algorithm 1 Offline-training process of the DoubleDQN algorithm

Initialize parameters of Q network (ω) and target network (ω').
for episode = 1 to M **do**:
  Initialize $s_1 = \left[1, E_{ESS}(1), P_0^{peak}, \mu_{grid}(1), Q_d(1)\right]$.
  for t = 1 to T **do**:
    Select action a_t at given s_t using the greedy policy.
    Execute a_t in the CCHP environment and transit to the next state s_{t+1}.
    Get reward r_t.
    Store the experience (s_t, a_t, r_t, s_{t+1}) in the experience replay buffer.
    Extract a mini-batch of experience $(s_t^i, a_t^i, r_t^i, s_{t+1}^i)$ with the size N from the experience replay buffer.
    Calculate the loss function: $L(\omega) = E[r_t + \tau \cdot Q(s_{t+1}, argmax_a Q(s_{t+1}, a; \omega); \omega') - Q(s_t, a_t; \omega))]^2$.
    Update the weights of the Q network: $\omega = \omega - \psi \cdot \nabla_\omega L(\omega)$.
    Copy the weights ω into the target network every fixed number of time steps: $\omega' = \omega$
  end for
end for

Algorithm 2 Decision-making procedure of the proposed method

Input: Environment state observation s_t of time step t.
Output: Dispatch decision a_t for energy supply units.
Load the weights ω of the Q network trained by Algorithm 1.
for time step = 1 to T **do**:
  Select action $a_t = \pi(s_t; \omega)$.
  Execute a_t in the CCHP environment and transit to the next state s_{t+1}.
  Get reward r_t.
end for

4. Case Study

In this section, the effectiveness and superiority of the proposed method are verified by comparison with the designed DQN-based method and benchmark policies. In addition, the proposed method is further tested under extended scenarios.

4.1. Simulation Setup

In order to evaluate the proposed DoubleDQN-based EDCS method, a CCHP system located in EXPO Site (Shanghai, China) is taken as a subject for the case study, the structure of which is shown in Figure 1. The system provides cooling for 28 office buildings on the site, and their working hours of them are from 6:00 to 18:00 on weekdays. This paper uses the historical cooling load data of these buildings from 2018 to 2020 for the proposed method of training and testing. The cooling load data from May to July is used to train, and the cooling load data from August is used to test. The length of the dispatch period is 720 h (that is, a full month). In addition, considering the buildings' working hours, 19:00 on the previous month's last day and 18:00 on the current month's last day are regarded as the start and end indexes of the dispatch period, respectively.

The CCHP system parameters are provided in Table 1. Additionally, the demand charge unit price μ_{demand}, the natural gas price μ_{gas} and the selling electricity price μ_{sell} are 42 RMB/kW, 2.57 RMB/m^3, and 0.568 RMB/kWh, respectively. The purchasing electricity price μ_{grid} is the time-of-use price: the valley tariff is 0.232 RMB/kWh (22:00~5:00), the peak tariff is 1.062 RMB/kWh (8:00~10:00, 13:00~14:00 and 18:00~20:00), and the flat tariff is 0.716 RMB/kWh at all other times.

The hyperparameters of DoubleDQN are shown in Table 2. Meanwhile, two common DRL algorithms DQN and dueling deep Q network (DuelingDQN) are introduced for use in subsequent subsections. Their hyperparameters are consistent with those of DoubleDQN.

Table 1. CCHP system parameters.

Parameter	Value	Parameter	Value
η_{ICE}	0.46	$\alpha_{EC}, \beta_{EC}, \gamma_{EC}$	0.013, 474.293, 1.615
$n^{ICE}_{max}, n^{EC}_{max}, n^{AC}_{max}$	2, 4, 2	$\alpha_{ICE}, \beta_{ICE}, \gamma_{ICE}$	−1.021, 1312.225, 0.005
$P^{ICE}_{rated}, Q^{EC}_{rated}$ (kW)	1600, 4700	$\alpha_{ST}, \beta_{ST}, \gamma_{ST}$	0.005, 0.062, 2.970
Q^{ST}_{max} (kWh)	10,000	COP_{EC}	5.2
E^{ST}_{rated} (kWh)	70,000	COP_{AC}	1.3

Table 2. Hyperparameters of DoubleDQN.

Description	Training Value	Description	Training Value
Size of input	5	Mini-batch size	128
No. of hidden layers	3	Discount factor	0.925
Size of each hidden layer	128, 512, 128	Learning rate	0.0005
Size of output	2	Weights of reward	λ_1: 5×10^{-5}, λ_2: 6×10^{-4}, θ: 0.002, ε_{peak}: 3300
Activation function for each hidden layer	ReLU	Optimizer	Adam

The proposed algorithm is implemented using the deep learning framework PyTorch 1.8.0, and the simulation experiments are carried out on a computer equipped with AMD Ryzen7 5700U CPU and 16 G RAM.

4.2. Off-Line Training Process

Figure 5 shows the cumulative reward performance of DoubleDQN, DQN, and DuelingDQN during the offline training process, where the full line represents the average. In the beginning, since the DRL agent is unfamiliar with the environment, the selected action results in a large variation in cumulative reward. As the training process continues, the agent begins to optimize the strategy to accumulate greater rewards. After about 32 episodes, DuelingDQN starts to converge, which is slightly prior to DoubleDQN (45 episodes) and DQN (68 episodes). However, DoubleDQN eventually obtains the greatest cumulative reward at convergence, which is far higher than that of DuelingDQN. This shows DoubleDQN's superior learning performance in exploring the optimal strategy. Additionally, due to DuelingDQN's poor cumulative reward performance, it is not considered in subsequent sections.

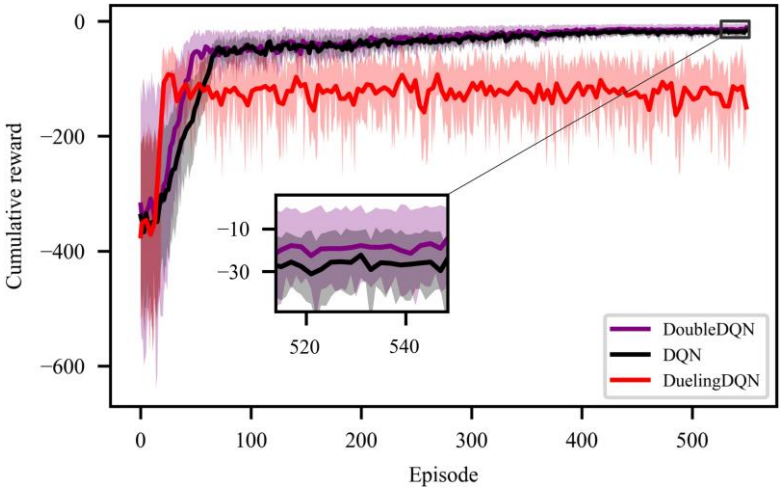

Figure 5. Cumulative reward performance of DoubleDQN, DQN, and DuelingDQN.

4.3. Dispatch Result Evaluation

Well-trained DRL agents from both the DoubleDQN and DQN are run during the test month (August 2020) to evaluate their dispatch results. Additionally, two benchmark policies are designed for comparison. The benchmark policies are described as follows: (a) Rule-based policy: during the valley electricity price period, ECs are given the priority to providing cooling, followed by ST and ACs, and the priority is completely reversed at all other times; (b) Shortsighted policy (based on DRL): the reward function defined in Equation (27) only consists of the energy cost and cooling error, and DRL agents are constructed using DQN and DoubleDQN, respectively.

The dispatch results and computational time of each method above are shown in Table 3, where the cooling error is expressed by the ratio of the total error to the total load, and the shortsighted policy is identified by the letter "a".

Table 3. Dispatch results and computational time of each method.

Method	Total Cost (RMB)	Energy Cost (RMB)	Demand Charge (RMB)	Rate of Cooling Error (%)	Total Online Running Time (s)	Total Offline Training Time (s)
Rule-based policy	1,678,249	1,524,251	153,998	0	20.27	-
DoubleDQN	1,152,830	1,013,176	139,654	0.312	1.44	973
DoubleDQN-a	1,256,845	995,455	261,390	0.608	1.46	995
DQN	1,154,421	1,011,638	142,783	0.550	1.41	1204
DQN-a	1,263,219	993,489	262,730	0.652	1.53	1013

As observed from Table 3, the three DRL methods have a significant advantage in on-line running time. Compared with the rule-based policy, the average time for DoubleDQN, DoubleDQN-a, DQN, and DQN-a to generate a set of dispatch decisions is reduced by 92.89%, 92.79%, 93.04%, and 92.45%, respectively. That's because the rule-based policy has to take a few seconds at each time step to make a series of logical judgments based on different environment information, so as to output the dispatch strategy. On the contrary, the DRL-based methods can make decisions in less than 4.25 ms by using the DNN that is well trained during the off-line training phase, which is far less than the dispatch interval of the rule-based policy. So, they can meet the requirements for real-time operation better.

It can be also found that the rule-based policy outperforms all DRL-based methods in controlling the cooling error. However, since it can only rigidly make the dispatch strategy according to the electricity price signal, the highest total cost is obtained as a consequence. On the other hand, compared to the rule-based policy, DoubleDQN-a saves 34.69% in energy cost while DoubleDQN saves only 32.53%, and a similar difference holds comparing DQN-a to DQN. It suggests that the shortsighted policy can better save energy costs and is more suitable for scenes without demand charges. However, DoubleDQN shows a greater advantage in the demand charge, which is 2.19~46.57% lower than other methods, thus obtaining the lowest total intra-month cost. Additionally, DoubleDQN also controls the cooling error at a relatively low level, indicating the demand-side user's comfort zone is well preserved.

To explore DoubleDQN's advantage in the demand charge, we compare the electricity dispatch strategies of DoubleDQN, DoubleDQN-a, and rule-based policy for a typical week of the test month, as shown in Figure 6. As observed, under DoubleDQN-a's dispatch strategy, the grid mainly supplies the electricity load, and the electric power purchase in the flat electricity price period is far higher than other methods, especially reaching a peak of 5739 kW in 70 h. However, under DoubleDQN's dispatch strategy, the peak electric power purchase is reduced from 3578 kW (rule-based policy) and 5739 kW (DoubleDQN-a) to 3112 kW by increasing ICEs' total electric power output, which greatly reduces the demand charge and flattens the electric power purchase curve. Additionally, part of the electricity load under this strategy is shifted to the nighttime when both the cooling load and electricity price are relatively low, which saves energy costs and maintains electricity

stability. Therefore, through the above comparison, it can be easily found that DoubleDQN exhibits better capabilities of demand response and peak demand management.

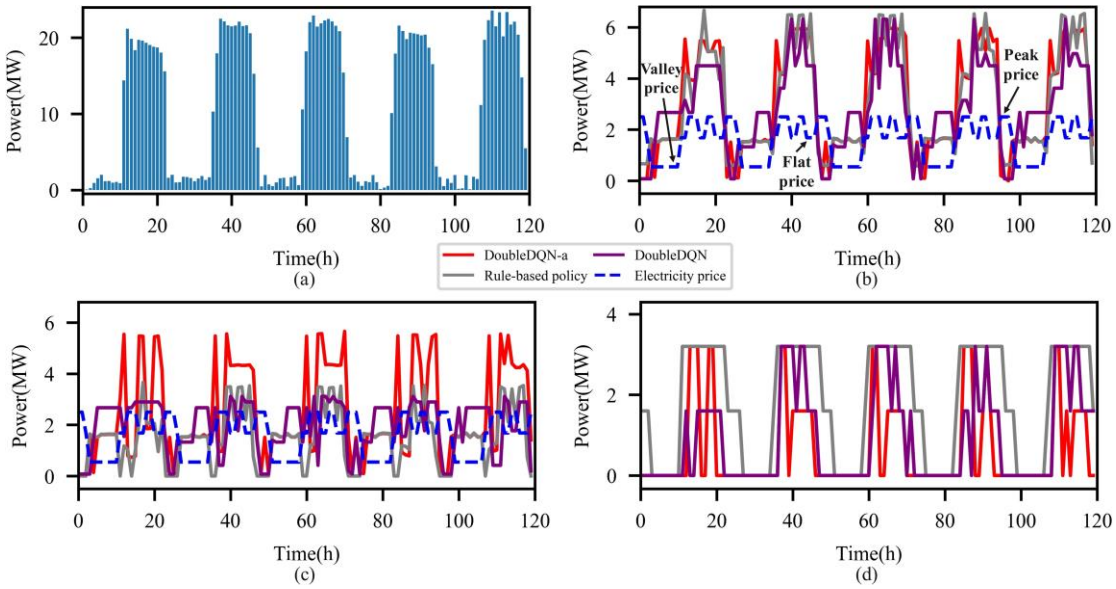

Figure 6. Electricity dispatch strategies of three methods for a typical week: (**a**) cooling load (**b**) electricity load (**c**) electric power purchase (**d**) electric power generation.

Figure 7 further illustrated the dispatch strategy of DoubleDQN for a typical day of the test month. It can be observed that the supply and demand of electric and cooling power are balanced throughout the day. As the electricity purchasing price is in the valley period (19:00~5:00), ECs consume the electric power purchased from the grid for providing cooling, and the redundant energy is stored in ST to release when requiring more cooling (6:00~17:00). When the cooling load and electricity price are both relatively high (9:00~16:00), ECs maintain the high total cooling output, while ICEs consume the natural gas for electric power generation to curb the peak electric power purchase, and the waste heat is used to produce cool through ACs in the corresponding period. This shows that DoubleDQN can flexibly manage the output of multiple units according to the environment information, so as to achieve the optimal energy dispatch.

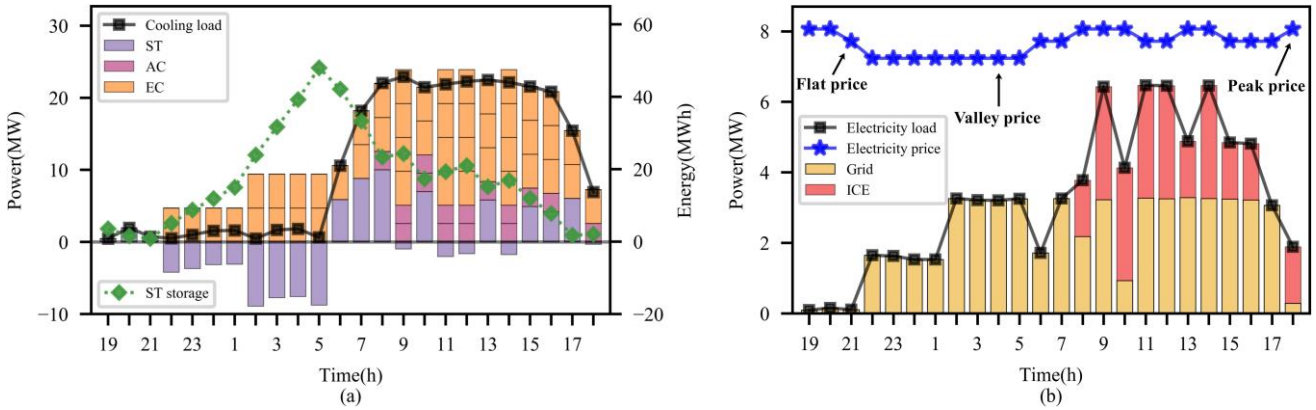

Figure 7. Dispatch strategy of DoubleDQN for a typical day: (**a**) cooling load balance (**b**) electricity load balance.

4.4. Extending the Proposed Method to Different Scenarios

In this subsection, different scenarios are designed to further test the proposed DoubleDQN-based method, and the test month is still applied. The scenarios are:

Scenario C_1: The proposed method is tested in new CCHP system models generated with different system parameters to test its generalization.

Scenario C_2: The cooling load at each time step t randomly fluctuates in the range of $[1-\alpha\%, 1+\alpha\%]$ to test the proposed method's robustness to the uncertain load, where $\alpha \in \{5,\ 10,\ 15,\ 20\}$.

Scenario C_3: Three typical unit failures are designed to test the proposed method's effectiveness in handling sudden unit failure. Those three-unit failures occur randomly at each time step, and the probability distribution of which is shown in Table 4, where the value pair (A, B) represents the maximum runnable number of ECs and ICEs at time step t, respectively.

Table 4. Probability distribution of unit failures.

Unit Failure (A, B)	Probability (%)
(4, 1)	35
(3, 2)	15
(3, 1)	5

4.4.1. Scenario C_1

Ten new CCHP system models are generated with different system parameters, and the variation of which follows a normal distribution $N(\zeta,\ 0.1\zeta)$, where ζ is the raw parameter. The dispatch results for these 10 models under DoubleDQN, DQN and rule-based policy are compared in Figure 8. It can be observed that similar to the results in Table 3, the rule-based policy obtains the lowest cooling error, while the highest energy cost and demand charge are obtained at the same time; DQN obtains the lowest energy cost and highest cooling error. In contrast, DoubleDQN can properly balance the above three objectives, and thus stably bringing the lowest total cost and relatively lower cooling error to all CCHP system models. Therefore, it can be easily concluded that DoubleDQN can adapt to unseen physical environments after being trained offline in a fixed environment.

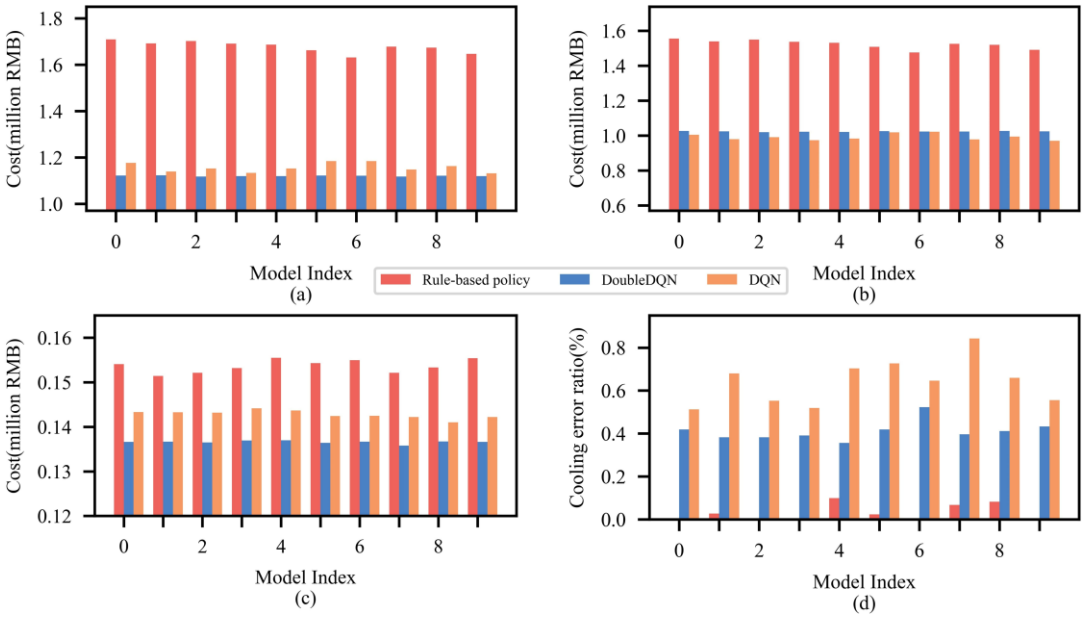

Figure 8. Comparison of dispatch results from DoubleDQN, DQN and rule-based policy: (**a**) total cost (**b**) energy cost (**c**) demand charge (**d**) cooling error ratio.

4.4.2. Scenario C_2

For each load fluctuation degree α, 150 experiments are carried out respectively, and the MPC-based method, which is widely used to solve uncertainty problems, is introduced

for the comparison. MPC-based method: (a) making dispatch decisions based on the rolling load-prediction, and the selected rolling horizon is set as 4 h; (b) the load-prediction error follows a normal distribution $N(0,\ 0.2\sigma)$, where σ is the actual load value.

As shown in Figure 9, the dispatch results of DoubleDQN, DQN and MPC in a total of 600 experiments are plotted and compared by violin plot. It can be observed that with the increase of load fluctuation degree α, the three methods' distributions of total cost and cooling error become more divergent. However, compared with DQN and MPC, the results of DoubleDQN show no significant discretization and achieve the lowest mean instead. Especially when the load fluctuation degree α is 5%, the superiority of DoubleDQN is more prominent. This shows that, unlike the MPC-based method, which relies on load-forecast accuracy, DoubleDQN efficiently deals with the uncertain load by directly making dispatch decisions based on real-time observations, and thus provides the dispatch strategy with higher stability and economic benefits. So, the proposed method can be relatively easier applied in real-world scenarios, especially when the prediction information is noisy or even missing.

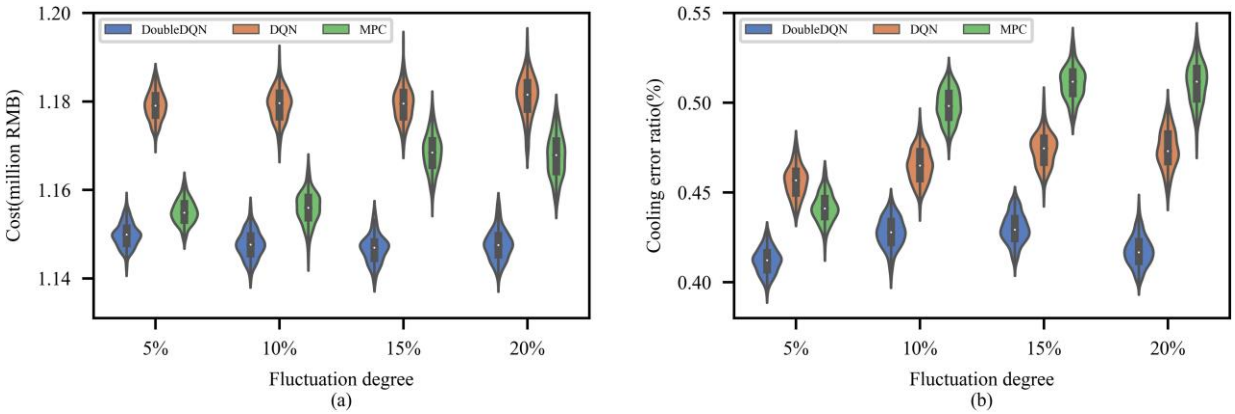

Figure 9. Comparison of dispatch results from DoubleDQN, DQN, and MPC: (**a**) total cost (**b**) cooling error ratio.

4.4.3. Scenario C_3

For the unit failures defined in Table 4, the corresponding DRL agents are constructed using DoubleDQN and DQN respectively to be called on-demand during the operation period of CCHP system. The dispatch results of DoubleDQN, DQN and rule-based policy in 150 experiments are compared in Table 5.

Table 5. Dispatch results of each method.

Method	Total Cost (RMB)			Ratio of Cooling Error (%)		
	Min	Mean	Max	Min	Mean	Max
DoubleDQN	**1,148,908**	**1,152,980**	**1,158,866**	**0.292**	**0.412**	**0.531**
DQN	1,168,379	1,185,324	1,189,573	0.358	0.659	1.031
Rule-based policy	1,644,612	1,683,547	1,691,356	0.014	0.305	0.804

It can be seen from the table that the average total cost of DoubleDQN is 2.72% and 31.51% lower than that of DQN and the rule-based policy, respectively. Meanwhile, DoubleDQN also achieves better performance than these methods in terms of minimum and maximum, maintaining relatively high economic benefits. On the other hand, compared with DQN, DoubleDQN has obtained a performance closer to that of the rule-based policy in controlling cooling error, well preserving the comfort zone for the demand-side user.

Figure 10 further compares the dispatch strategies of DoubleDQN for a typical day both under normal and faulty conditions, i.e., sudden unit failures. Under the faulty condition, DoubleDQN increases the energy storage of ST by 42% by increasing the cool production of ECs during the nighttime. Therefore, the higher storage allows ST to release more cooling energy than the normal condition, so as to reduce the cool production of ECs and ACs by 11% and 19% during the daytime, respectively. On the other hand, the electricity load shows a trend of shifting from the daytime to the nighttime accordingly, which reduces the purchasing electricity cost and the daytime ICE electricity generation. Therefore, it can be easily concluded that by rationally planning the energy storage and release of ST, DoubleDQN reduces the CCHP system's dependence degree of both ECs and ICEs while balancing the supply and demand, so as to efficiently handle unpredictable unit failures during the operation period, which can meet the requirements for practical application.

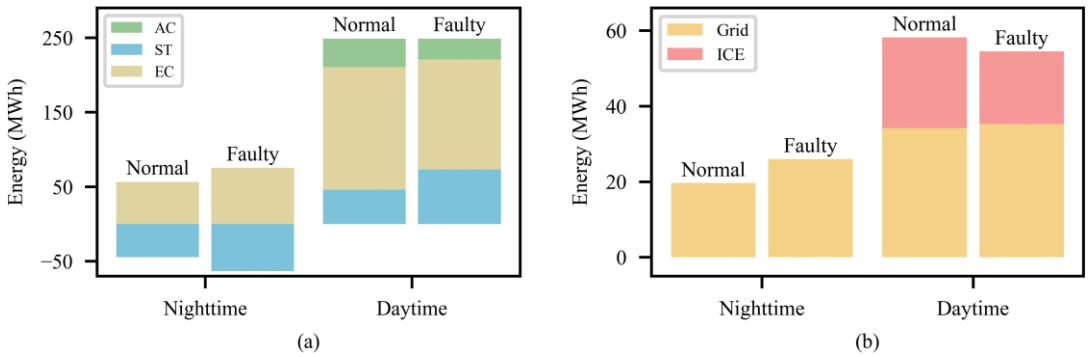

Figure 10. Dispatch strategies of DoubleDQN both under normal and faulty conditions: (**a**) normal condition (**b**) faulty condition.

5. Conclusions

This paper focuses on the summer energy dispatching problem of the CCHP system. Aiming at minimizing the total intra-month cost and balancing the supply and demand of energy, a model-free DoubleDQN-based method is proposed to generate an optimal dispatch strategy. Different from the traditional method, this method makes dispatch decisions directly based on the real-time observed electricity price and cooling load, avoiding the suboptimal dispatching problem caused by prediction error. Through the simulation results, the following conclusions can be drawn:

(1) Compared with DRL algorithms DQN and DuelingDQN, DoubleDQN shows better learning performance during off-line training and obtains the greatest cumulative reward at convergence.

(2) The proposed method shows good demand response and peak shift ability. By restraining the peak electric power purchase of the CCHP system to below 3112 kW, the total intra-month cost is further reduced by 0.13~31.32% compared with the designed DRL methods and the rule-based policy through greater demand charge advantage. In addition, the method also considers the decision speed and thermal comfort, which not only meets the requirements of real-time operation but also well preserves the comfort zone for the demand-side user.

(3) In dealing with unknown system parameters, load uncertainty and sudden unit failure, the proposed method provides the dispatch strategy with higher stability and economic benefits for the CCHP system, showing strong generalization and potential for application in real scenarios.

On the other hand, the future study work will focus on two directions: one is to focus on the improvement of the traditional DRL algorithm to reduce the time and computing resources in DRL training; the other is to include environmental pollution into optimization indicators.

Author Contributions: Conceptualization, W.G. and Y.L.; methodology, Y.L.; software, Y.L.; formal analysis, W.G.; writing-review and editing, W.G. and Y.L.; visualization, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Shanghai Municipal Science and Technology Commission (grant number 18040501800), and the National Natural Science Foundation of China (grant number 51706129).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy.

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

Nomenclature

AC	absorption chiller
AE	auxiliary equipment
CCHP	combined cooling, heating and power
COP	coefficient of performance
CT	cooling tower
DC	demand charge
DRL	Deep reinforcement learning
EC	electric chiller
ECT	energy cost
GB	gas boiler
ICE	internal combustion engine
ST	storage tank
a_t	selected action of agent at time step t
c_t	cost at time step t, RMB
C_{total}	total cost, RMB
E^{ST}	remaining energy storage of ST, kW
n^{AC}	number of running ACs
n^{EC}	number of running ECs
n^{ICE}	number of running ICEs
P_d	electricity load, kW
P_e	electric power consumption of supply unit, kW
$P_{e,a}$	allocated electric power consumption of supply unit, kW
$P_{e,i}^{EC}$	electric power consumption of ith running EC, kW
P^{grid}	exchanging electric power with the grid, kW
P_i^{ICE}	electric power output of ith running ICE, kW
P_{rated}^{ICE}	rated electric power generation, kW
P_t^{peak}	peak electric power purchase in last t time steps, kW
Q_d	cooling load, kW
Q_i^{AC}	cooling output of ith running AC, kW
Q_i^{EC}	cooling output of ith running EC, kW
Q_{rated}^{EC}	rated cooling generation of EC, kW
Q^{ST}	storing/releasing power of ST, kW
Q^{waste}	low-grade waste heat, kW
$Q_\pi(s,a)$	state-action value function
r_t	received reward at time step t
s_t	environment state at time step t
V_i^{ICE}	gas consumption of ith running ICE, m^3/h
α, β, γ	electric power consumption coefficients
ΔQ_d	difference between supply and demand of cooling power, kW
ε_{LHV}	low calorific value of natural gas, kWh/m^3
ε_{peak}	threshold of peak electric power purchase, kW
η_{ICE}	electric efficiency of ICE
θ	extra penalty of reward function
λ	weighting factors of reward function
μ_{demand}	unit price of demand charge, RMB/kW
μ_{gas}	natural gas price, RMB/m^3
μ_{grid}	purchasing electricity price, RMB/kWh
μ_{sell}	selling electricity price, RMB/kWh
τ	discount factor
ψ	learning factor
ω/ω'	weight parameters of Q network/target network

References

1. Zhang, D.; Zhang, B.; Zheng, Y.; Zhang, R.; Liu, P.; An, Z. Economic assessment and regional adaptability analysis of CCHP system coupled with biomass-gas based on year-round performance. *Sustain. Energy Technol. Assess.* **2021**, *45*, 101141. [CrossRef]
2. Jalili, M.; Ghasempour, R.; Ahmadi, M.H.; Chitsaz, A.; Holagh, S.G. An integrated CCHP system based on biomass and natural gas co-firing: Exergetic and thermo-economic assessments in the framework of energy nexus. *Energy Nexus* **2022**, *5*, 100016. [CrossRef]
3. Gu, W.; Lu, S.; Wu, Z.; Zhang, X.; Zhou, J.; Zhao, B.; Wang, J. Residential CCHP microgrid with load aggregator: Operation mode, pricing strategy, and optimal dispatch. *Appl. Energy* **2017**, *205*, 173–186. [CrossRef]
4. Li, L.; Mu, H.; Gao, W.; Li, M. Optimization and analysis of CCHP system based on energy loads coupling of residential and office buildings. *Appl. Energy* **2014**, *136*, 206–216. [CrossRef]
5. Wang, X.; Xu, Y.; Fu, Z.; Guo, J.; Bao, Z.; Li, W.; Zhu, Y. A dynamic interactive optimization model of CCHP system involving demand-side and supply-side impacts of climate change. Part II: Application to a hospital in Shanghai, China. *Energy Convers. Manag.* **2022**, *252*, 115139. [CrossRef]

6. Saberi, K.; Pashaei-Didani, H.; Nourollahi, R.; Zare, K.; Nojavan, S. Optimal performance of CCHP based microgrid considering environmental issue in the presence of real time demand response. *Sustain. Cities Soc.* **2019**, *45*, 596–606. [CrossRef]
7. Lu, S.; Gu, W.; Zhou, J.; Zhang, X.; Wu, C. Coordinated dispatch of multi-energy system with district heating network: Modeling and solution strategy. *Energy* **2018**, *152*, 358–370. [CrossRef]
8. Shan, J.; Lu, R. Multi-objective economic optimization scheduling of CCHP micro-grid based on improved bee colony algorithm considering the selection of hybrid energy storage system. *Energy Rep.* **2021**, *7*, 326–341. [CrossRef]
9. Kang, L.; Yuan, X.; Sun, K.; Zhang, X.; Zhao, J.; Deng, S.; Liu, W.; Wang, Y. Feed-forward active operation optimization for CCHP system considering thermal load forecasting. *Energy* **2022**, *254*, 124234. [CrossRef]
10. Ghersi, D.E.; Amoura, M.; Loubar, K.; Desideri, U.; Tazerout, M. Multi-objective optimization of CCHP system with hybrid chiller under new electric load following operation strategy. *Energy* **2021**, *219*, 119574. [CrossRef]
11. Li, Y.; Tian, R.; Wei, M.; Xu, F.; Zheng, S.; Song, P.; Yang, B. An improved operation strategy for CCHP system based on high-speed railways station case study. *Energy Convers. Manag.* **2020**, *216*, 112936. [CrossRef]
12. Wang, J.; Yang, Y. A hybrid operating strategy of combined cooling, heating and power system for multiple demands considering domestic hot water preferentially: A case study. *Energy* **2017**, *122*, 444–457. [CrossRef]
13. Lin, H.; Yang, C.; Xu, X. A new optimization model of CCHP system based on genetic algorithm. *Sustain. Cities Soc.* **2020**, *52*, 101811. [CrossRef]
14. Zhu, G.; Chow, T.-T. Design optimization and two-stage control strategy on combined cooling, heating and power system. *Energy Convers. Manag.* **2019**, *199*, 111869. [CrossRef]
15. Ma, D.; Zhang, L.; Sun, B. An interval scheduling method for the CCHP system containing renewable energy sources based on model predictive control. *Energy* **2021**, *236*, 121418. [CrossRef]
16. Hu, K.; Wang, B.; Cao, S.; Li, W.; Wang, L. A novel model predictive control strategy for multi-time scale optimal scheduling of integrated energy system. *Energy Rep.* **2022**, *8*, 7420–7433. [CrossRef]
17. Majidi, M.; Mohammadi-Ivatloo, B.; Anvari-Moghaddam, A. Optimal robust operation of combined heat and power systems with demand response programs. *Appl. Therm. Eng.* **2019**, *149*, 1359–1369. [CrossRef]
18. Siqin, Z.; Niu, D.; Wang, X.; Zhen, H.; Li, M.; Wang, J. A two-stage distributionally robust optimization model for P2G-CCHP microgrid considering uncertainty and carbon emission. *Energy* **2022**, *260*, 124796. [CrossRef]
19. Cheng, Z.; Jia, D.; Li, Z.; Si, J.; Xu, S. Multi-time scale dynamic robust optimal scheduling of CCHP microgrid based on rolling optimization. *Int. J. Electr. Power Energy Syst.* **2022**, *139*, 107957. [CrossRef]
20. Batista Abikarram, J.; McConky, K.; Proano, R. Energy cost minimization for unrelated parallel machine scheduling under real time and demand charge pricing. *J. Clean. Prod.* **2019**, *208*, 232–242. [CrossRef]
21. van Zoest, V.; El Gohary, F.; Ngai, E.C.H.; Bartusch, C. Demand charges and user flexibility—Exploring differences in electricity consumer types and load patterns within the Swedish commercial sector. *Appl. Energy* **2021**, *302*, 117543. [CrossRef]
22. Hledik, R. Rediscovering Residential Demand Charges. *Electr. J.* **2014**, *27*, 82–96. [CrossRef]
23. Zhang, Y.; Augenbroe, G. Optimal demand charge reduction for commercial buildings through a combination of efficiency and flexibility measures. *Appl. Energy* **2018**, *221*, 180–194. [CrossRef]
24. Maldonado-Ramirez, A.; Rios-Cabrera, R.; Lopez-Juarez, I. A visual path-following learning approach for industrial robots using DRL. *Robot. Comput. Manuf.* **2021**, *71*, 102130. [CrossRef]
25. Fuchs, A.; Heider, Y.; Wang, K.; Sun, W.; Kaliske, M. DNN2: A hyper-parameter reinforcement learning game for self-design of neural network based elasto-plastic constitutive descriptions. *Comput. Struct.* **2021**, *249*, 106505. [CrossRef]
26. Guo, Y.; Ma, J. DRL-TP3: A learning and control framework for signalized intersections with mixed connected automated traffic. *Transp. Res. Part C: Emerg. Technol.* **2021**, *132*, 103416. [CrossRef]
27. Alabdullah, M.H.; Abido, M.A. Microgrid energy management using deep Q-network reinforcement learning. *Alex. Eng. J.* **2022**, *61*, 9069–9078. [CrossRef]
28. Gao, G.; Li, J.; Wen, Y. DeepComfort: Energy-Efficient Thermal Comfort Control in Buildings Via Reinforcement Learning. *IEEE Internet Things J.* **2020**, *7*, 8472–8484. [CrossRef]
29. Yang, T.; Zhao, L.; Li, W.; Zomaya, A.Y. Dynamic energy dispatch strategy for integrated energy system based on improved deep reinforcement learning. *Energy* **2021**, *235*, 121377. [CrossRef]
30. Hasanvand, S.; Rafiei, M.; Gheisarnejad, M.; Khooban, M.-H. Reliable Power Scheduling of an Emission-Free Ship: Multiobjective Deep Reinforcement Learning. *IEEE Trans. Transp. Electrification* **2020**, *6*, 832–843. [CrossRef]
31. Guo, C.; Wang, X.; Zheng, Y.; Zhang, F. Real-time optimal energy management of microgrid with uncertainties based on deep reinforcement learning. *Energy* **2022**, *238*, 121873. [CrossRef]
32. Zhou, S.; Hu, Z.; Gu, W.; Jiang, M.; Chen, M.; Hong, Q.; Booth, C. Combined heat and power system intelligent economic dispatch: A deep reinforcement learning approach. *Int. J. Electr. Power Energy Syst.* **2020**, *120*, 106016. [CrossRef]
33. Ma, J.; Qin, J.; Salsbury, T.; Xu, P. Demand reduction in building energy systems based on economic model predictive control. *Chem. Eng. Sci.* **2012**, *67*, 92–100. [CrossRef]
34. Jiang, Z.; Risbeck, M.J.; Ramamurti, V.; Murugesan, S.; Amores, J.; Zhang, C.; Lee, Y.M.; Drees, K.H. Building HVAC control with reinforcement learning for reduction of energy cost and demand charge. *Energy Build.* **2021**, *239*, 110833. [CrossRef]
35. Yang, T.; Zhao, L.; Li, W.; Wu, J.; Zomaya, A.Y. Towards healthy and cost-effective indoor environment management in smart homes: A deep reinforcement learning approach. *Appl. Energy* **2021**, *300*, 117335. [CrossRef]

36. Tan, H.; Zhang, H.; Peng, J.; Jiang, Z.; Wu, Y. Energy management of hybrid electric bus based on deep reinforcement learning in continuous state and action space. *Energy Convers. Manag.* **2019**, *195*, 548–560. [CrossRef]
37. Du, Y.; Zandi, H.; Kotevska, O.; Kurte, K.; Munk, J.; Amasyali, K.; Mckee, E.; Li, F. Intelligent multi-zone residential HVAC control strategy based on deep reinforcement learning. *Appl. Energy* **2021**, *281*, 116117. [CrossRef]
38. Carta, S.; Ferreira, A.; Podda, A.S.; Recupero, D.R.; Sanna, A. Multi-DQN: An ensemble of Deep Q-learning agents for stock market forecasting. *Expert Syst. Appl.* **2021**, *164*, 113820. [CrossRef]
39. Tong, Z.; Ye, F.; Liu, B.; Cai, J.; Mei, J. DDQN-TS: A novel bi-objective intelligent scheduling algorithm in the cloud environment. *Neurocomputing* **2021**, *455*, 419–430. [CrossRef]
40. Ding, Y.; Ma, L.; Ma, J.; Suo, M.; Tao, L.; Cheng, Y.; Lu, C. Intelligent fault diagnosis for rotating machinery using deep Q-network based health state classification: A deep reinforcement learning approach. *Adv. Eng. Informatics* **2019**, *42*, 100977. [CrossRef]
41. Han, S.; Zhou, W.; Lu, J.; Liu, J.; Lü, S. NROWAN-DQN: A stable noisy network with noise reduction and online weight adjustment for exploration. *Expert Syst. Appl.* **2022**, *203*. [CrossRef]
42. Park, S.; Yoo, Y.; Pyo, C.-W. Applying DQN solutions in fog-based vehicular networks: Scheduling, caching, and collision control. *Veh. Commun.* **2022**, *33*, 100397. [CrossRef]
43. Zhang, W.; Gai, J.; Zhang, Z.; Tang, L.; Liao, Q.; Ding, Y. Double-DQN based path smoothing and tracking control method for robotic vehicle navigation. *Comput. Electron. Agric.* **2019**, *166*. [CrossRef]
44. Dong, P.; Chen, Z.-M.; Liao, X.-W.; Yu, W. A deep reinforcement learning (DRL) based approach for well-testing interpretation to evaluate reservoir parameters. *Pet. Sci.* **2022**, *19*, 264–278. [CrossRef]