# Low-carbon Economic Dispatch of Electricity-Heat-Gas Integrated Energy Systems Based on Deep Reinforcement Learning

Yuxian Zhang, Yi Han, Deyang Liu, and Xiao Dong

***Abstract*—The optimal dispatch methods of integrated energy systems (IESs) currently struggle to address the uncertainties resulting from renewable energy generation and energy demand. Moreover, the increasing intensity of the greenhouse effect renders the reduction of IES carbon emissions a priority. To address these issues, a deep reinforcement learning (DRL)-based method is proposed to optimize the low-carbon economic dispatch model of an electricity-heat-gas IES. In the DRL framework, the optimal dispatch model of the IES is formulated as a Markov decision process (MDP). A reward function based on the reward-penalty ladder-type carbon trading mechanism (RPLT-CTM) is introduced to enable the DRL agents to learn more effective dispatch strategies. Moreover, a distributed proximal policy optimization (DPPO) algorithm, which is a novel policy-based DRL algorithm, is employed to train the DRL agents. The multithreaded architecture enhances the exploration ability of the DRL agents in complex environments. Experimental results illustrate that the proposed DPPO-based IES dispatch method can mitigate carbon emissions and reduce the total economic cost. The RPLT-CTM-based reward function outperforms the CTM-based methods, providing a 4.42% and 6.41% decrease in operating cost and carbon emission, respectively. Furthermore, the superiority and computational efficiency of DPPO compared with other DRL-based methods are demonstrated by a decrease of more than 1.53% and 3.23% in the operating cost and carbon emissions of the IES, respectively.**

***Index Terms*—Integrated energy system (IES), carbon trading, optimal dispatch, deep reinforcement learning (DRL), distributed proximal policy optimization.**

## I. Introduction

THE limitations of traditional energy sources and the diversity of human needs pose considerable challenges to current energy structures [1]. Integrated energy systems (IESs) can optimize the overall energy utilization while exploiting renewable energy sources. Therefore, IESs are considered as key elements in the development of future human society [2], [3]. In contrast to traditional separated energy systems, IES enables the comprehensive management and economic dispatch (ED) of multiple energy resources, thus improving the complementary utilization of electricity, heat, gas, and transportation [4].

Recently, research work on the ED of IESs has received increasing attention. However, the fluctuation and randomness of renewable energy and load represent a source of uncertainty, thus complicating the solution to the ED problem for IESs [5]. As an important branch of machine learning, deep reinforcement learning (DRL) has the advantage of self-learning through interactive trial and error in a dynamic environment [6]. DRL has been applied to solve sequential decision-making problems with uncertainties [7]. Hence, DRL appears to be suitable for renewable energy and electric system optimization problems, which involve complex nonlinearities and uncertainties [8].

A relevant aspect to consider in the development of IESs is global warming, which is caused by the emission of greenhouse gases with $CO_2$ as the main component [9]. Reducing $CO_2$ emissions has become a major goal in the development of IESs. The carbon trading mechanism (CTM) is an essential market mechanism that guides energy companies to meet emission targets. The CTM has attracted increasing international attention, leading to the development of a framework for an international carbon market, which was proposed at the 26$^{th}$ United Nations Climate Change Conference [10]. For example, the Hainan International Carbon Emission Trading Center in China completed its first cross-border carbon emission trading in January 2023 [11]. The impact of the CTM on the optimal scheduling problem of low-carbon IESs requires further study and discussion.

Traditional dispatch methods are based on day-ahead forecasting information. However, these methods do not consider uncertainties of load demand and renewable energy generation. Mathematical programming-based methods have been developed to solve ED problems while considering these uncertainties. Reference [12] proposes a scenario-based stochastic optimization (SO) method for IESs to address the uncertainties in energy demand and renewable generation. Reference [13] proposes a robust optimization (RO)-based day-

Manuscript received: October 15, 2022; revised: February 8, 2023; accepted: April 5, 2023. Date of CrossCheck: April 5, 2023. Date of online publication: May 1, 2023.

This work was supported in part by the National Natural Science Foundation of China (No. 61102124).



Y. Zhang (corresponding author), Y. Han, and D. Liu are with the School of Electrical Engineering, Shenyang University of Technology, Shenyang, China, and D. Liu is also with the College of Electrical Engineering, Yingkou Institute of Technology, Yingkou, China (e-mail: y.x.zhang@sut.edu.cn; hanyi_2020@163.com; ldy1988@outlook.com).

X. Dong is with Beijing Ke Dong Co., Ltd., NARI Group Corporation, Beijing, China (e-mail: 732880355@qq.com).

DOI: 10.35833/MPCE.2022.000671

ahead dispatch model that considers the effects of outdoor temperature uncertainty on thermal comfort. Reference [14] proposes an RO-based energy management framework for the optimal day-ahead dispatch of a multi-energy microgrid accounting for uncertainties of the power market price. Reference [15] proposes a hybrid SO-RO method for the coordinate scheduling of a multi-energy system, in which erratic and high-risk wind power production is modeled by RO, whereas energy demands with a detectible probability distribution are modeled as stochastic scenarios. The hybrid RO-SO method in [16] can model uncertain variables with different characteristics separately by combining the advantages of the SO and RO methods. However, this method requires the design of an optimal dispatch framework according to the specific properties of the random variables involved, while considering the operating cost and reliability of the system. The distributionally robust optimization (DRO) [17] method has gradually gained attention because it obtains decision results by considering the worst probability distribution of uncertain parameters.

However, these dispatch methods have certain limitations. Scenario-based SO may require the generation of several scenarios based on probability distributions, resulting in a severe increase in computational burden. More importantly, the optimal dispatch results may not satisfy the constraints of scenarios that are not considered [18]. As the RO-based method may attempt to avoid the impact of uncertainties on system operation, its results can be too conservative and are often not conducive to the economical operation of IESs [19]. The hybrid SO-RO method cannot overcome these disadvantages. Although the DRO-based method combines the advantages of SO and RO, it requires complex modeling and solving processes.

Control theory-based methods such as model predictive control (MPC) have also been used to address uncertainties in the optimal operation problem. Reference [20] proposes an MPC-based bi-level optimal integration scheme for the space heating load of buildings to achieve the economical and reliable scheduling of the heating system in the presence of uncertainties. Information gap decision theory (IGDT) [21] is another method for addressing uncertainties. In [22], a multi-objective IGDT-based method is applied to handle the uncertainties associated with wind and photovoltaic (PV) power predictions. Although MPC-based methods use receding horizon optimization to offset uncertainties, they still employ renewable energy generation predictions. Furthermore, the selection of some parameters in the IGDT method is operator-dependent. A summary of the advantages and disadvantages of the aforementioned methods for solving dispatch problems with uncertainties is presented in Table I.

TABLE I
ADVANTAGES AND DISADVANTAGES OF METHODS FOR SOLVING DISPATCH PROBLEMS WITH UNCERTAINTIES

| Reference | Method | Description |
|---|---|---|
| [12] | SO | Many scenarios need to be generated. A severe computational burden may be incurred. The optimal dispatch results may not satisfy the constraints of scenarios that are not considered. |
| [13], [14] | RO | The results are conservative because the worst case of uncertainty is considered. |
| [15], [16] | SO-RO | The operating cost and reliability of the system are considered. Appropriate scenarios are required. |
| [17] | DRO | The advantages of SO and RO are combined. The modeling and solving processes are complex. |
| [20] | MPC | Rolling optimization is applied to offset uncertainty. The process is complicated, and the optimization quality relies on the forecast accuracy of uncertain variables. |
| [21], [22] | IGDT | The choice of some coefficients is subjective. |
| [23]-[45] | DRL | Instead of relying on prior knowledge, the agent collects data by interacting with the environment and learning from data. The agent can be applied to real-time dispatch after offline training. |

In contrast to the aforementioned methods, the DRL agent collects data by interacting with the IES environment and learns a dispatch strategy from the data. In some studies, DRL algorithms have been applied in discrete action spaces to solve optimal dispatch problems that consider uncertainties in microgrids [23], home energy management [24], distributed energy systems [25], and multi-energy microgrids [26]. However, such a discrete action space not only affects the accuracy of the dispatch results, but also causes the dispatch strategy to lose flexibility. Some studies have applied DRL algorithms to solve optimal dispatch problems with a continuous action space. In [27], an online energy management system is built using policy gradient (PG) algorithm. Several other alternatives have been proposed to address the optimal scheduling problem of microgrids, including asynchronous advantage actor-critic (A3C) [28], deep deterministic policy gradient (DDPG) [29], and proximal policy optimization (PPO) [30]. However, all these studies consider only the electrical network as the research object. Therefore, further research on the advantages of multi-energy network coupling for optimal dispatch should be conducted.

In [31], a PPO-based renewable energy conversion strategy is applied to reduce the operating costs of an IES. To solve the ED problem of a combined heat and power (CHP) system, [32] adopts a distributed PPO-based method. Reference [33] proposes an improved DDPG algorithm for the optimal scheduling of an electricity-heat IES. Reference [34] develops a real-time autonomous energy management strategy for a residential multi-energy system based on DDPG. Reference [35] proposes a PPO-based joint load scheduling strategy to reduce the energy costs of a household multi-energy system. In [36], a DDPG-based dynamic energy conversion and management strategy is used to coordinate economic costs and peak load shifting targets. Reference [37] develops an optimal dispatch framework based on A3C to handle the dynamic changes on the supply and demand sides of an

IES. However, these studies do not adequately discuss the methods for reducing the carbon emissions of the system; in fact, they only employ the operating cost of the system as the dispatching target.

To satisfy the energy demands of an IES and minimize operating costs and pollutant emissions, [38] proposes a DRL-based intelligent energy management system. However, it can only be applied to discrete action space. Reference [39] employs the soft actor-critic (SAC) algorithm to solve the optimal dispatch problem of electricity-gas IES using economical operation and low carbon emissions as the objectives of the dispatch model. In [40], an SAC-based energy dispatch strategy is developed to optimize the multiple objectives of an IES, including minimizing operational costs and realizing economical low-carbon operation.

Reference [41] designs a multi-agent cooperative control framework for the energy management of a multi-energy hub using an attention mechanism based on multi-agent deep reinforcement learning (MADRL). Moreover, in [42], MADRL is employed to solve the optimal dispatch problem of an IES considering energy trading, and in [41] and [42], the carbon emission target is added to the reward function as a penalty term. However, the CTM is not considered. Thus, the IES cannot profit from the sale of carbon rights.

Several studies have attempted to introduce the CTM into DRL-based frameworks. Reference [43] proposes a model-free safe DRL method for the real-time automatic optimal energy management of a renewable-based energy hub with various energy components, in which both the system energy cost and carbon emissions are minimized. In [44], an IES co-trading market that includes electricity, natural gas, and CTM is proposed. The coordinative optimization problem associated to energy management is solved using an improved multi-agent DDPG algorithm. In [45], a joint peer-to-peer energy and carbon allowance trading mechanism for a building community is proposed, considering both the flexibility of local trading and decarbonization of building multi-energy systems. In these studies, the combination of the CTM and the low-carbon ED problem of IES or the integrated energy trading market based on the CTM has demonstrated better results in controlling and reducing carbon emissions. However, as DRL is applied to solve such problems, the effectiveness of the CTM in helping agents learn low-carbon dispatch strategies should be discussed in detail. The incentive and penalty mechanism of the CTM for companies to reduce emissions is similar to the idea of designing a reward function for DRL. Therefore, this deserves to be discussed in depth, rather than being simply combined.

Most studies applying DRL methods to solve the optimal dispatch problem while accounting for uncertainties have not considered the carbon emissions of the system. Only a few studies have considered carbon emissions by introducing a traditional CTM-based reward function to obtain a low-carbon ED model for the IES. However, as the reward function affects the effectiveness of the strategy learned by the agent, it should be carefully designed within the DRL framework. Moreover, the introduction of CTM increases the complexity of the DRL environment. Hence, a more efficient algorithm is required for the agent to learn low-carbon ED strategies.

To address the existing research gap, a DRL-based dynamic energy dispatch method is proposed for the low-carbon economic operation of an electricity-heat-gas IES. A comparison of the elements considered in the development of our model and those presented in the reviewed models is presented in Table II.

TABLE II
COMPARISON BETWEEN PROPOSED MODEL AND REVIEWED MODELS

| Reference | Action space | | Energy | | | Dispatch | | CTM | |
|---|---|---|---|---|---|---|---|---|---|
| | Discrete | Continuous | Electricity | Heat | Gas | Economy | Emission | Traditional | Ladder-type |
| [23]-[25] | √ | | √ | | | √ | | | |
| [26] | √ | | √ | √ | | √ | | | |
| [27]-[30] | | √ | √ | | | √ | | | |
| [31]-[34] | | √ | √ | √ | | √ | | | |
| [35] | √ | √ | √ | | √ | √ | | | |
| [36], [37] | | √ | √ | √ | √ | √ | | | |
| [38] | √ | | √ | √ | | √ | √ | | |
| [39] | | √ | √ | | √ | √ | √ | | |
| [40] | | √ | √ | √ | √ | √ | √ | | |
| [41] | | √ | √ | | √ | √ | √ | | |
| [42] | | √ | √ | √ | | √ | √ | | |
| [43]-[45] | | √ | √ | √ | | √ | √ | √ | |
| Proposed model | | √ | √ | √ | √ | √ | √ | | √ |

To achieve low-carbon operation of the system, a reward-penalty ladder-type CTM (RPLT-CTM) is introduced into the DRL framework. The RPLT-CTM models the principles that guide enterprises to reduce emissions. For this reason, we decide to use the RPLT-CTM-based reward function with variable carbon trading prices to guide the agent more effectively in learning the low-carbon economic scheduling strategy for the IES. Moreover, to solve the optimal scheduling prob-

lem, the distributed proximal policy gradient (DPPO) algorithm is introduced, which is a policy-based DRL algorithm that is less sensitive to hyperparameters and can avoid large policy updates with undesirable action selections.

The major contributions can be summarized as follows.

1) A DRL-based method for low-carbon ED of an electricity-heat-gas IES, which considers economics and carbon emissions, is established. The low-carbon ED is mathematically modeled as a Markov decision process (MDP).

2) The RPLT-CTM is introduced into the DRL framework to realize low-carbon ED. Compared with the traditional CTM, the RPLT-CTM-based reward function has been proven to guide the DRL agent in formulating an improved low-carbon ED strategy.

3) To address the increased complexity introduced by the low-carbon objective, the DPPO algorithm with a distributed architecture is introduced to train the DRL agent. A comparative analysis demonstrates the computational effectiveness and superiority of this algorithm.

The remainder of this paper is organized as follows. Section II presents the electricity-heat-gas IES, including the carbon trading cost calculation model for the RPLT-CTM-based IES, and the mathematical model for IES optimal dispatch. In Section III, the optimal dispatch problem is formulated as an MDP, and the DPPO-based method for IES optimal dispatch is described in detail. Simulation results and the corresponding analysis are presented in Section IV. Conclusions and future work are discussed in Section V.

## II. Electricity-Heat-Gas IES

The primary goal of the optimal dispatch of the IES is to improve the economic benefits of the system, i. e., on the premise of satisfying the energy demand, the output of each piece of equipment at each time step is effectively arranged to achieve the optimal economic operation. Furthermore, to realize low-carbon operation of the system, the RPLT-CTM is introduced to incorporate carbon trading costs into the operating costs of the system. To this end, we establish a comprehensive ED model that considers the RPLT-CTM. The structure of electricity-heat-gas IES is shown in Fig. 1.

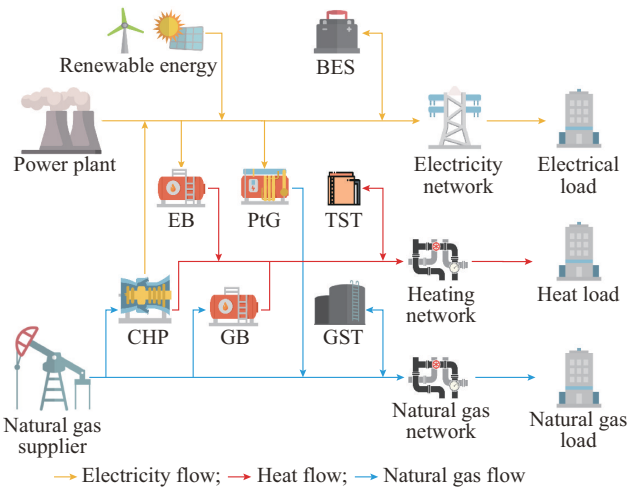


Fig. 1. Structure of electricity-heat-gas IES.

The IES consists of energy suppliers, renewable energy generation devices, load demand, coupling devices, and energy storage devices. Renewable energy generation devices include wind turbines (WTs) and PV generators. The load demand includes electrical, heat, and gas loads. The coupling equipment includes a CHP, power-to-gas (PtG), and gas boiler (GB). The energy storage equipment includes battery energy storage (BES), gas storage tanks (GSTs), and thermal storage tanks (TSTs).

### A. Carbon Trading Cost Calculation Model for RPLT-CTM-based IES

The CTM can guide energy companies to reduce emissions, and its essence is to treat carbon credit allowances as freely tradable commodities [46]. The specific model is presented as follows.

#### 1) Initial Carbon Credit Allocation Model

The allocation of initial carbon credits is a prerequisite for low-carbon power dispatch. The initial carbon emission allowance allocation is performed using the free allocation method.

In the IES model, the electricity purchased from the external grid is produced by coal-fired units. In addition to the equipment in the IES that generates carbon emissions, natural gas loads are also considered. The CHP unit is considered as heat supply equipment, and its carbon credits are allocated according to the equivalent total heat supply. Thus, the power generated by the CHP units needs to be converted into an equivalent heat supply. The model is expressed as:

$$\begin{cases} E_{IES,c}=E_{grid,c}+E_{CHP,c}+E_{GB,c}+E_{gload,c} \\ E_{grid,c}=\lambda_e\sum_{t=1}^{T}p_{grid}(t)\Delta t \\ E_{CHP,c}=\lambda_h\sum_{t=1}^{T}\left(\varphi p_{CHP}(t)+h_{CHP}(t)\right)\Delta t \\ E_{GB,c}=\lambda_h\sum_{t=1}^{T}h_{GB}(t)\Delta t \\ E_{gload,c}=\lambda_{gas}\sum_{t=1}^{T}q_{load}(t)\Delta t \end{cases} \tag{1}$$

where $E_{IES,c}$ is the total carbon credit allowance of the IES; $E_{grid,c}$, $E_{CHP,c}$, and $E_{GB,c}$ are the carbon credit allowances for coal-fired units, CHP, and GB, respectively; $E_{gload,c}$ is the carbon credit allowance received by the user for the consumption of natural gas; $\Delta t$ is the interval for each time step; $p_{grid}(t)$, $p_{CHP}(t)$, $h_{CHP}(t)$, and $h_{GB}(t)$ are the output power of the coal-fired units, CHP, and GB at time step $t$, respectively; $q_{load}(t)$ is the flow rate of the natural gas load at time step $t$; $\lambda_e$, $\lambda_h$, and $\lambda_{gas}$ are the carbon credit allocation factors for the electricity supply equipment, heat supply equipment, and natural gas load, respectively; and $\varphi$ is the conversion factor of power generation into heat supply, which is taken as 6 MJ/kWh.

#### 2) Carbon Emission Calculation Model

In the IES, the operation of the CHP units and GB generates carbon emission. The electricity purchased from the ex-

ternal grid comes from coal-fired units, the operation of which generates carbon emissions. The consumption of natural gas loads, mainly through combustion, also generates carbon emissions. The working process of the PtG unit involves the absorption of $CO_2$. The carbon emission model of the IES is:

$$\begin{cases} E_{IES,e}=E_{grid,e}+E_{CHP,e}+E_{GB,e}+E_{gload,e}-E_{PtG,e} \\ E_{grid,e}=\beta_e\sum_{t=1}^{T}p_{grid}(t)\Delta t \\ E_{CHP,e}=\beta_h\sum_{t=1}^{T}\left(\varphi p_{CHP}(t)+h_{CHP}(t)\right)\Delta t \\ E_{GB,e}=\beta_h\sum_{t=1}^{T}h_{GB}(t)\Delta t \\ E_{gload,e}=\beta_{gas}\sum_{t=1}^{T}q_{load}(t)\Delta t \\ E_{PtG,e}=\beta_{PtG}\sum_{t=1}^{T}p_{PtG}(t)\Delta t \end{cases} \tag{2}$$

where $E_{IES,e}$ is the total carbon emission of the IES; $E_{grid,e}$, $E_{CHP,e}$, $E_{GB,e}$, and $E_{gload,e}$ are the carbon emissions generated the by coal-fired units, CHP, GB, and natural gas load, respectively; $E_{PtG,e}$ is the amount of $CO_2$ absorbed in the energy conversion process of the PtG unit; $\beta_e$, $\beta_h$, and $\beta_{gas}$ are the carbon emission factors for the electricity supply equipment, heat supply equipment, and natural gas load, respectively; $p_{PtG}(t)$ is the electric power consumed by the PtG unit at time step $t$; and $\beta_{PtG}$ is the parameter for the absorption of $CO_2$ in the energy conversion of the PtG unit.

*3) Carbon Trading Model*

The RPLT-CTM [47] divides several net carbon emission intervals and guides the system in reducing $CO_2$ emissions through incentives and penalties. In addition, the carbon trading price shows a stepwise increase with the cumulative carbon trading volume, as shown in Fig. 2.

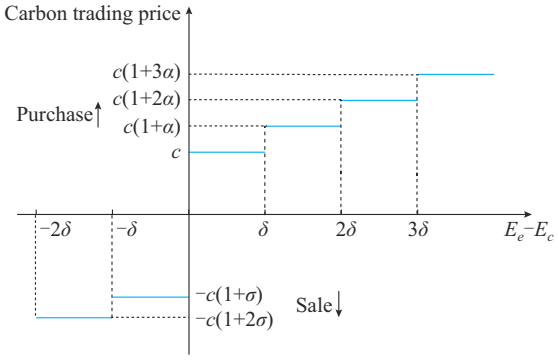


Fig. 2. Relationship between carbon trading price and cumulative carbon trading volume.

The mathematical model of the reward and penalty ladder-type carbon trading is expressed as:

$$E_{IES}(t)=E_{IES,e}(t)-E_{IES,c}(t) \tag{3}$$

$$E_{IES}=E_{IES,e}-E_{IES,c} \tag{4}$$

$$C_{CT}(t)=\begin{cases} c(1+2\sigma)E_{IES}(t) & E_{IES}\le-\delta \\ c(1+\sigma)E_{IES}(t) & -\delta<E_{IES}\le 0 \\ cE_{IES}(t) & 0<E_{IES}\le\delta \\ c(1+\alpha)E_{IES}(t) & 0<E_{IES}\le\delta \\ c(1+2\alpha)E_{IES}(t) & \delta<E_{IES}\le 2\delta \\ c(1+3\alpha)E_{IES}(t) & 3\delta<E_{IES} \end{cases} \tag{5}$$

where $E_{IES}(t)$ is the amount of carbon trading at time step $t$; $E_{IES}$ is the cumulative carbon trading volume; $C_{CT}(t)$ is the carbon trading cost of the IES at time step $t$; $c$ is the carbon trading price; $\alpha$ is the penalty factor, which is taken as 0.2; $\sigma$ is the reward factor, which is taken as 0.25; and $\delta$ is the length of the carbon trading range.

### *B. Mathematical Model for IES Optimal Dispatch*

*1) Objective Function*

The primary goal of the IES dynamic energy dispatch is to improve the economy and environmental friendliness of the system while meeting the constraints. The objective function is mainly composed of energy purchase and carbon trading costs. The objective function $F$ of the optimal dispatch is defined as:

$$F=\min\sum_{t=1}^{T}\left(C_E(t)+C_{CT}(t)\right) \tag{6}$$

where $C_E(t)$ is the energy purchase cost at time step $t$.

*2) Cost of Energy Purchase*

To satisfy the electricity-heat-gas load demand, the system purchases energy from energy suppliers as fuel for the operation of the coupled equipment. The equipment that consumes electrical energy includes the PtG units and electric boiler (EB), and the equipment that consumes natural gas is the CHP units and GB. This cost is expressed as:

$$C_E(t)=C_{power}(t)+C_{gas}(t) \tag{7}$$

$$C_{power}(t)=\varepsilon_e(t)\,p_{grid}(t)\Delta t \tag{8}$$

$$C_{gas}(t)=\varepsilon_{gas}(t)\,q_{gas}(t)\Delta t \tag{9}$$

where $C_{power}(t)$ and $C_{gas}(t)$ are the costs of the purchased electricity and natural gas, respectively; $q_{gas}(t)$ is the output flow rate of the natural gas supplier; $\varepsilon_e(t)$ is the electricity price; and $\varepsilon_{gas}(t)$ is the natural gas price.

*3) Constraints*

The constraints of IES dynamic scheduling consist of energy balance, equipment operation, and energy supplier constraints.

1) Energy balance constraints

To meet the electricity-heat-gas load demand at each time step, the energy balance constraints are:

$$p_{grid}(t)+p_{RE}(t)+p_{CHP}(t)+p_{BES}(t)= p_{load}(t)+p_{EB}(t)+p_{PtG}(t) \tag{10}$$

$$h_{CHP}(t)+h_{EB}(t)+h_{GB}(t)+h_{TST}(t)=h_{load}(t) \tag{11}$$

$$q_{gas}(t)+q_{PtG}(t)+q_{GST}(t)=q_{load}(t)+q_{CHP}(t)+q_{GB}(t) \tag{12}$$

where $p_{RE}(t)$ is the renewable energy generation; $p_{BES}(t)$ is the charging/discharging power of the BES; $p_{EB}(t)$ is the

electric power consumed by the EB; $h_{EB}(t)$ is the power output of the EB; $h_{TST}(t)$ is the charging/discharging power of the TST; $q_{PtG}(t)$ is the output flow rate of PtG; $q_{GST}(t)$ is the charging/discharging power of the GST; $q_{CHP}(t)$ is the flow rate of natural gas consumed by CHP; $q_{GB}(t)$ is the flow rate of natural gas consumed by the GB; and $p_{load}(t)$ and $h_{load}(t)$ are the electrical load and heat load, respectively.

2) Equipment operation constraints

① Energy supply devices

a) CHP

The CHP unit provides heat and electricity to the system and acts as an energy provider in the electricity and heating networks. The mathematical model of the CHP unit is expressed as:

$$p_{CHP}(t) = k_{CHP} h_{CHP}(t) \tag{13}$$

$$q_{CHP}(t) = \frac{p_{CHP}(t) + h_{CHP}(t)}{\eta_{CHP} H_{GV}} \tag{14}$$

where $k_{CHP}$ is the thermoelectric ratio of CHP; $\eta_{CHP}$ is the efficiency of CHP; and $H_{GV}$ is the high calorific value of natural gas, which is taken as 39 MJ/m$^3$.

The power output and ramping rate constraints of the CHP unit are given by (15)-(18).

$$p_{CHP}^{\min} \le p_{CHP}(t) \le p_{CHP}^{\max} \tag{15}$$

$$h_{CHP}^{\min} \le h_{CHP}(t) \le h_{CHP}^{\max} \tag{16}$$

$$-R_{CHP}^{\text{down}} \Delta t \le p_{CHP}(t) - p_{CHP}(t-1) \le R_{CHP}^{\text{up}} \Delta t \tag{17}$$

$$-R_{CHP}^{\text{down}} \Delta t \le h_{CHP}(t) - h_{CHP}(t-1) \le R_{CHP}^{\text{up}} \Delta t \tag{18}$$

where $p_{CHP}^{\min}$ and $p_{CHP}^{\max}$ are the lower and upper bounds of the output electric power, respectively; $h_{CHP}^{\min}$ and $h_{CHP}^{\max}$ are the lower and upper bounds of the output heat power of CHP, respectively; $p_{CHP}(t-1)$ and $h_{CHP}(t-1)$ are the output electric and heat power of CHP at time step $t-1$, respectively; and $R_{CHP}^{down}$ and $R_{CHP}^{up}$ are the ramping rates of CHP.

b) PtG

The PtG unit converts electric power into gas. The relationship between the electric power consumption and the natural gas supply is expressed as:

$$q_{PtG}(t) = \frac{\eta_{PtG} p_{PtG}(t)}{H_{GV}} \tag{19}$$

where $\eta_{PtG}$ is the efficiency of PtG.

The power and ramping rate constraints of the PtG unit are shown in (20) and (21), respectively.

$$p_{PtG}^{\min} \le p_{PtG}(t) \le p_{PtG}^{\max} \tag{20}$$

$$-R_{PtG}^{\text{down}} \Delta t \le p_{PtG}(t) - p_{PtG}(t-1) \le R_{PtG}^{\text{up}} \Delta t \tag{21}$$

where $p_{PtG}^{\min}$ and $p_{PtG}^{\min}$ are the lower and upper bounds of the consumed electric power, respectively; $p_{PtG}(t-1)$ is the electric power consumed by PtG at time step $t-1$; and $R_{PtG}^{down}$ and $R_{PtG}^{up}$ are the ramping rates of PtG.

c) EB

The EB converts electric power into heat to satisfy the heat load. The relationship between the electric power consumption and the heat supply is expressed as:

$$h_{EB}(t) = \eta_{EB} p_{EB}(t) \tag{22}$$

where $\eta_{EB}$ is the efficiency of the EB.

The power output and ramping rate constraints of the EB are shown in (23) and (24), respectively.

$$h_{EB}^{\min} \le h_{EB}(t) \le h_{EB}^{\max} \tag{23}$$

$$-R_{EB}^{\text{down}} \Delta t \le h_{EB}(t) - h_{EB}(t-1) \le R_{EB}^{\text{up}} \Delta t \tag{24}$$

where $h_{EB}^{\min}$ and $h_{EB}^{\max}$ are the lower and upper bounds of the output heat power of the EB, respectively; $h_{EB}(t-1)$ is the power output of the EB at time step $t-1$; and $R_{EB}^{down}$ and $R_{EB}^{up}$ are the ramping rates of the EB.

d) GB

The GB converts natural gas power into heat power, which is used to supplement the remaining heat load demand when the CHP heat supply is insufficient. The relationship between the natural gas power consumption and the heat supply is expressed as:

$$h_{GB}(t) = \eta_{GB} q_{GB}(t) H_{GV} \tag{25}$$

where $\eta_{GB}$ is the efficiency of the GB.

The power output and ramping rate constraints of the GB are given by (26) and (27), respectively.

$$h_{GB}^{\min} \le h_{GB}(t) \le h_{GB}^{\max} \tag{26}$$

$$-R_{GB}^{\text{down}} \Delta t \le h_{GB}(t) - h_{GB}(t-1) \le R_{GB}^{\text{up}} \Delta t \tag{27}$$

where $h_{GB}^{\min}$ and $h_{GB}^{\max}$ are the lower and upper bounds of the output heat power of the GB, respectively; $h_{GB}(t-1)$ is the power output of the GB at time step $t-1$; and $R_{GB}^{down}$ and $R_{GB}^{up}$ are the ramping rates of the GB.

② Energy storage equipment

a) BES

The BES can store excess energy in the system, which can be reasonably discharged to meet the electrical demand of customers in case of insufficient power supply. For the BES, the state of charge (SOC) is a key operational parameter that directly reflects the remaining energy of the device.

$$SOC_{\min} \le SOC(t) \le SOC_{\max} \tag{28}$$

$$SOC(t) = SOC(t-1) - \eta_{BES} \frac{p_{BES}(t)}{Q_{BES}} \Delta t \tag{29}$$

$$\eta_{BES} = \begin{cases} \eta_{ch} & p_{BES}(t) < 0 \\ 1/\eta_{dis} & p_{BES}(t) \ge 0 \end{cases} \tag{30}$$

where $SOC(t)$ and $SOC(t-1)$ are the SOCs of the BES at time steps $t$ and $t-1$, respectively; $SOC_{\min}$ and $SOC_{\max}$ are the lower and upper bounds of the SOC of the BES, respectively; $Q_{BES}$ is the capacity of the BES; $\eta_{BES}$ is the charging/discharging efficiency of the BES; and $\eta_{ch}$ and $\eta_{dis}$ are the charging and discharging coefficients, respectively.

b) TST

Similar to the BES, the TST can store excess heat and supply the heat needed for a heat load in the event of a heating shortage. Similar to the definition of SOC, the heat storage degree (HSD) is defined to monitor the amount of heat energy that can be stored in the equipment.

$$HSD_{\min} \le HSD(t) \le HSD_{\max} \tag{31}$$

$$HSD(t) = HSD(t-1) - \eta_{TST} \frac{h_{TST}(t)}{Q_{TST}} \Delta t \tag{32}$$

$$\eta_{TST}=\begin{cases}\eta_{ch} & h_{TST}(t)<0\\ 1/\eta_{dis} & h_{TST}(t)\geq 0\end{cases} \tag{33}$$

where $HSD(t)$ and $HSD(t-1)$ are the HSDs of the TST at time steps $t$ and $t-1$, respectively; $HSD_{\min}$ and $HSD_{\max}$ are the lower and upper bounds of the HSD of the TST, respectively; $Q_{TST}$ is the capacity of the TST; and $\eta_{TST}$ is the charging/discharging efficiency of the TST.

c) GST

The gas storage degree (GSD) of the GST is defined to monitor the amount of natural gas energy that can be stored in the equipment.

$$GSD_{\min}\leq GSD(t)\leq GSD_{\max} \tag{34}$$

$$GSD(t)=GSD(t-1)-\eta_{GST}\frac{q_{GST}(t)}{Q_{GST}}\Delta t \tag{35}$$

$$\eta_{GST}=\begin{cases}\eta_{ch} & q_{GST}(t)<0\\ 1/\eta_{dis} & q_{GST}(t)\geq 0\end{cases} \tag{36}$$

where $GSD(t)$ and $GSD(t-1)$ are the GSDs of the GST at time steps $t$ and $t-1$, respectively; $GSD_{\min}$ and $GSD_{\max}$ are the lower and upper bounds of the GSD of the GST, respectively; $Q_{GST}$ is the capacity of the GST; and $\eta_{GST}$ is the charging/discharging efficiency of the GST.

③ Energy supplier constraints

In the dispatching model established in this paper, electricity and natural gas need to be purchased from external sources to supply the equipment and meet the load demand. The energy supply device satisfies the following constraints.

$$p_{grid}^{\min}\leq p_{grid}(t)\leq p_{grid}^{\max} \tag{37}$$

$$q_{gas}^{\min}\leq q_{gas}(t)\leq q_{gas}^{\max} \tag{38}$$

where $p_{grid}^{\min}$ and $p_{grid}^{\max}$ are the lower and upper bounds of the output electric power of the coal-fired units, respectively; and $q_{gas}^{\min}$ and $q_{gas}^{\max}$ are the lower and upper bounds of the output gas flow rate of the supplier, respectively.

## III. DPPO-BASED METHOD FOR IES OPTIMAL DISPATCH

In this section, the IES optimal dispatch is formulated as an MDP, and the specific reinforcement learning algorithm is explained.

### A. RL Framework

MDP is a mathematically idealized form of the RL problem and a theoretical framework for achieving goals through interactive learning. An MDP consists of a state space $S$, action space $A$, state transition probability function $P$, reward function $R$, and discount coefficient $\gamma$.

An RL framework is built to solve the low-carbon ED problem for an IES, as shown in Fig. 3. In the RL environment for the IES dispatch problem, the state space includes information on the electric load, heat load, natural gas load, predicted value of renewable energy output, and state of the energy storage equipment. The action space includes the output power of the CHP units, electricity-to-gas equipment, EBs, and GBs as well as the power of electricity and natural gas purchased from external suppliers. The rewards include the optimization targets defined above such as operating costs and carbon trading costs. During the training process, the dispatch agent observes the load information and equipment states in the environment at time step $t$, adjusts the output power of each piece of equipment to satisfy the load demand, and then receives the reward and the next state $s_{t+1}$ from the environment back to the agent. The fundamental elements of the MDP can be formulated as follows.

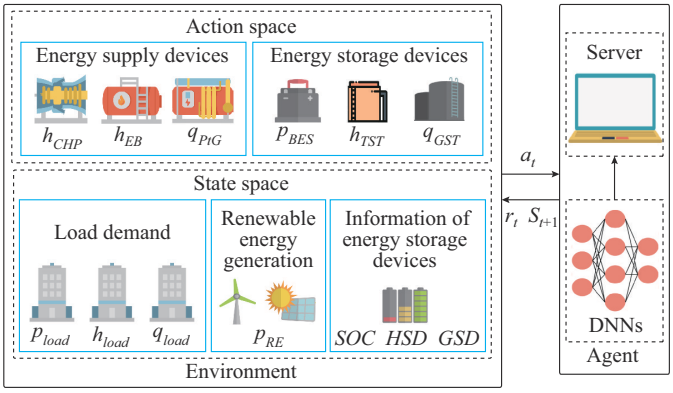


Fig. 3. RL framework for IES optimal dispatch.

#### 1) State Space

The state space $S$ contains the information that describes the state of the IES, and the dispatch agent decides the dispatch strategy based on the observed state at each time step. Specifically, the state space $S$ includes the electrical load $p_{load}(t)$, heat load $h_{load}(t)$, natural gas load $q_{load}(t)$, power output of renewable energy $p_{RE}(t)$, SOC of the BES $SOC(t)$, status (HSD) of the TST $HSD(t)$, and status (GSD) of the GST $GSD(t)$. Consequently, the state space is defined as:

$$\begin{aligned}s_t=\{&p_{load}(t),h_{load}(t),q_{load}(t),p_{RE}(t),\\ &SOC(t),HSD(t),GSD(t),t\}\end{aligned} \tag{39}$$

#### 2) Action Space

The dispatch agent realizes the optimal scheduling strategy for the IES by controlling the electric and heat power outputs of CHP ($p_{CHP}(t)$, $h_{CHP}(t)$), heat power output of the EB $h_{EB}(t)$, heat power output of the GB $h_{GB}(t)$, the gas power output of PtG $q_{PtG}(t)$, electric power purchased from the main grid $p_{grid}(t)$, natural gas power purchased from the natural gas supplier $q_{gas}(t)$, electric power output of the BES $p_{BES}(t)$, heat power output of the TST $h_{TST}(t)$, and natural gas power output of the GST $q_{GST}(t)$. The electric and natural gas power consumed by each device in the system such as $q_{chp}(t)$ is calculated from its output power. The energies purchased from external energy suppliers, $p_{grid}(t)$ and $q_{gas}(t)$, are calculated using electric power balance constraints and gas power balance constraints, respectively. The heat power output of the GB $h_{GB}(t)$ can also be calculated using the heat power balance constraint. That is, when $h_{CHP}(t)$, $h_{EB}(t)$, $q_{PtG}(t)$, $p_{BES}(t)$, $h_{TST}(t)$, and $q_{GST}(t)$ are jointly determined, the other variables can be obtained immediately. Therefore, action space is expressed as:

$$a_t=\{h_{CHP}(t),h_{EB}(t),q_{PtG}(t),p_{BES}(t),h_{TST}(t),q_{GST}(t)\} \tag{40}$$

*3) Reward Function*

The reward function calculates the reward value $r_t$ based on the current state and action $(s_t, a_t)$, then returns it to the agent. The purpose of the reward is to guide the agent to accomplish the stated goal, i.e., low carbon emissions and ED of the IES. Therefore, the reward function includes the operating cost $C_E$ and carbon trading cost $C_{CT}$ of the system. Considering that the goal of the training agent in reinforcement learning is to maximize the cumulative reward, the reward value needs to be set to be a negative value. To accelerate convergence, a baseline $b$ is added to the reward function so that positive and negative reward values can be given. The reward function can be defined as:

$$r_t = -\left(C_E(s_t, a_t) + C_{CT}(s_t, a_t) - b\right) \tag{41}$$

where $b$ is taken as 30.

*4) Uncertainty of RL Environment*

The stochastic nature of renewable energy generation and multiple energy loads needs to be considered in the IES optimal dispatch problem. To enable the agent to handle this uncertainty, the RL environment for the optimal scheduling problem needs to be established with stochasticity. Before the start of training for each episode, the environment randomly samples the load data that satisfy the upper and lower bound limits.

In each episode, a group of states is generated within the upper and lower limits. The energy loads and the renewable energy generation are generated randomly within the predefined range, which means that the dispatch strategy given by the agent can handle not only the uncertainty of loads but also the uncertainty of renewable energy generation.

## B. DPPO

The DRL algorithm is introduced to solve the optimal dispatch problem for a continuous action space. PPO [48] is a policy-based DRL algorithm for solving continuous action decisions, which is proposed by Google's DeepMind team [49] based on PPO, drawing on the parallel training idea of A3C. DPPO is better suited for rich simulation environments that consider uncertainty. We introduce the DPPO algorithm to solve the problem of optimal dispatch of IES considering uncertainty. The equations for DPPO in this subsection can be found in [43] and [44].

The PPO algorithm is a policy-based DRL algorithm with an actor-critic architecture. The advantage function $A_\pi(s_t, a_t)$ is introduced to evaluate the goodness of action $a_t$ in state $s_t$.

$$A_\pi(s_t, a_t) = Q_\pi(s_t, a_t) - V_\pi(s_t) \tag{42}$$

The action-value ($Q$-value) function $Q_\pi(s_t, a_t)$ is used to evaluate the performance of policy $\pi$, and is defined as:

$$Q_\pi(s_t, a_t) = E_{(s_{t'}, a_{t'}) \sim \pi_\theta}\left[\sum_{t=0}^{\infty} \gamma^t r_t | S_t = s_t, A_t = a_t\right] \tag{43}$$

where $\pi_\theta$ is the policy $\pi$ with parameter $\theta$; and $\gamma$ is the reward discount factor.

The state-value function $V_\pi(s_t)$ is used to evaluate the quality of state $s_t$, and is expressed as:

$$V_\pi(s_t) = E_{a_t \sim \pi_\theta(\cdot|s_t)}\left[\sum_{t=0}^{\infty} \gamma^t r_t | S_t = s_t\right] \tag{44}$$

From (43) and (44), the value of the action value function $Q_\pi(s_t, a_t)$ represents the expectation of the cumulative reward for choosing action $a_t$ in state $s_t$ under the guidance of policy network $\pi$. Furthermore, the value of the state-value function $V_\pi(s_t)$ represents the expectation of the cumulative reward for all actions in state $s_t$ under policy $\pi$.

With the introduction of the advantage function $A_\pi(s_t, a_t)$, the original objective function can be rewritten as:

$$J^{\theta^\mu}(\theta^\mu) = E_{(s_t, a_t) \sim \pi_{\theta^{\mu'}}}\left[\frac{\pi_{\theta^\mu}(a_t|s_t)}{\pi_{\theta^{\mu'}}(a_t|s_t)} A_\pi^{\theta^{\mu'}}(s_t, a_t)\right] \tag{45}$$

where $\theta^\mu$ is the parameter of the policy network to be optimized; and $\theta^{\mu'}$ is the parameter of the policy network that interacts with the environment to sample data. This is the surrogate objective function.

Next, the clipped surrogate objective method is employed. The surrogate objective function is written as:

$$J_{PPO\text{-}Clip}^{\theta^{\mu'}}(\theta^\mu) = E_{(s_t, a_t) \sim \pi_{\theta^{\mu'}}}\Big[\min\left(\rho_\theta A^{\theta^{\mu'}}(s_t, a_t),\right.$$
$$\left. clip\left(\rho_\theta, 1-\varepsilon, 1+\varepsilon\right) A^{\theta^{\mu'}}(s_t, a_t)\right)\Big] \tag{46}$$

$$clip\left(\rho_\theta, 1-\varepsilon, 1+\varepsilon\right) = \begin{cases} 1-\varepsilon & \rho(\theta) < 1-\varepsilon \\ 1+\varepsilon & \rho(\theta) > 1+\varepsilon \\ \rho(\theta) & \text{otherwise} \end{cases} \tag{47}$$

$$\rho_\theta = \frac{\pi_{\theta^\mu}(a_t|s_t)}{\pi_{\theta^{\mu'}}(a_t|s_t)} \tag{48}$$

where $\varepsilon$ is a surrogate objective function clipping rate applied to limit the change in policy.

The clip function limits the probability ratio to a certain range and takes the maximum or minimum value if it is out of range. By clipping the probability ratio, changes in policy are maintained within a reasonable range. This ensures that the change in policy is not too intense when the advantage is positive and that the update direction is correct when the advantage is negative. Finally, the PPO algorithm updates the policy network parameters using gradient ascent.

$$\theta^\mu = \theta^\mu + \alpha^\mu \nabla_{\theta^\mu} J_{PPO\text{-}Clip}^{\theta^{\mu'}}(\theta^\mu) \tag{49}$$

where $\alpha^\mu$ is the learning rate of the policy network.

The PPO algorithm has an actor-critic architecture. After updating the policy network, i.e., actor network, the critic network is updated by minimizing the loss function based on temporal-different (TD) theory.

$$L(\theta^Q) = E_{a_t \sim \pi_{\theta^{\mu'}}(\cdot|s_t)}\left[\left(q_t - V(s_t)\right)^2\right] \tag{50}$$

$$q_t = r_t + \gamma V(s_{t+1}) \tag{51}$$

$$\theta^Q = \theta^Q + \alpha^Q \nabla_{\theta^Q} L(\theta^Q) \tag{52}$$

where $L(\theta^Q)$ is the loss function; and $\alpha^Q$ is the learning rate of the $Q$-value network, i.e., critic network.

To train the agent to obtain better performance in the established optimal IES scheduling environment, the agent must fully explore the environment to face different scenari-

os. Therefore, the PPO algorithm with distributed settings was introduced to achieve better training performance. DPPO includes workers and a chief, where the workers are set up as multiple threads responsible for interacting with their respective environments to sample data and provide the data to the chief for learning. All parallel threads share the same policy network parameters from the global learner. The chief updates the network parameters and passes the pre-updated parameters to the workers. Each worker does not compute or push the gradient of its own policy update to the chief; this method promotes the efficiency of the multithreaded data collection and reduces the difficulty in implementing the algorithm. The framework of the DPPO algorithm training process is illustrated in Fig. 4.

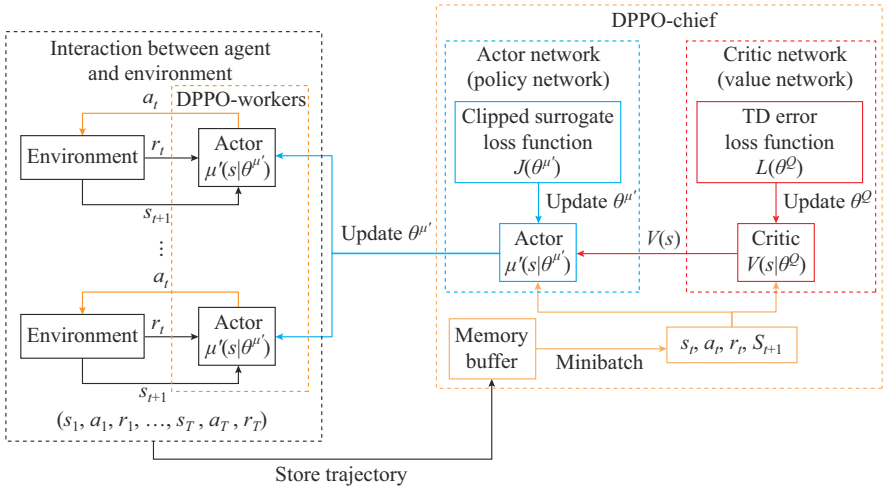


Fig. 4. Framework of DPPO algorithm training process.

The distributed setting of DPPO is reflected in the parallel collection of data based on the multithreaded worker network for the chief network update. In simple terms, DPPO can be understood as a multithreaded parallel PPO. The training process of DPPO is realized through multithreading and communication among multiple threads. The exploration thread of the workers and the update thread of the chief are not executed simultaneously and communicate through events. The flow of the alternating execution of multiple threads in DPPO is shown in Fig. 5.

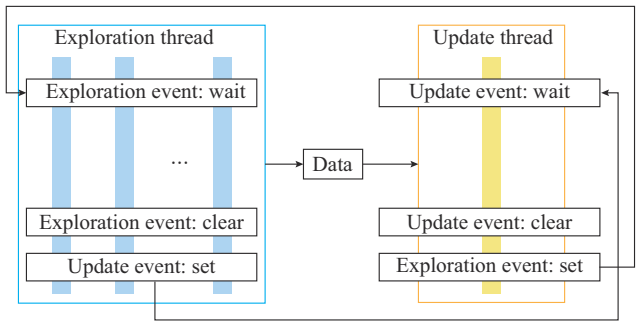


Fig. 5. Flow of alternating execution of multiple threads in DPPO.

At the beginning of training, the exploration event is set to be "set", and workers start interacting with the environment to collect data. The update event is set to be "clear" and enters the waiting state. In the exploration thread, the global variable *global_update_counter* is used to record the number of steps taken by the workers to interact with the environment. When the value of *global_update_counter* is larger than the mini-batch size, the update event is set to be "set" and the chief network starts to update. The exploration event is set to be "clear" and will enter the waiting state when running to "wait". After the chief network update is complete, the update event is set to be "clear" and suspended. The exploration event is set to be "set" and workers continue to interact with the environment to collect data. The offline training process of the DPPO algorithm is shown in the pseudocode in Algorithm 1.

## IV. Case Study

In this section, a platform for IES optimal scheduling is established and experiments are conducted using this IES platform to verify the superiority of the proposed DPPO-based dispatch method. The parameter settings, experimental details, and concluding analysis are presented in the following subsections.

### *A. Description of IES*

To demonstrate the performance of the proposed DPPO-based dispatch method, the IES shown in Fig. 1 is used as an example in the case study. The IES consists of a power grid, heating network, natural gas network, renewable generation devices, and energy storage equipment. In addition to using the CHP, GB, EB, and PtG to satisfy the load demand, energy can be purchased from external energy suppliers.

The purchasing electricity price is the time-of-use (TOU) price. The peak-time price is 12.3 ¢/kWh (12:00-20:00), the valley-time price is 4.2 ¢/kWh (00:00-08:00), and the flat-time price is 7.8 ¢/kWh at all other time. The natural gas price is fixed at 49 ¢/m$^3$. In the RPLT-CTM, the carbon trading price is 40 $/t, and the length of the carbon trading range is 2 t. The property parameters of RPLT-CTM including carbon credit allocation factors and carbon emission factors are listed in Table III.

**Algorithm 1**: off-line training process of DPPO

Initialize parameters $\theta^{\mu}$ and $\theta^{Q}$ randomly
Initialize old actor parameters: $\theta^{\mu'} \leftarrow \theta^{\mu}$
*exploration_event.set*(), *update_event.clear*()
*global_update_counter* = 0
**for** *episode* = 1 to *N* **do**
 **if not** *exploration_event.set*() **then**
  *exploration_event.wait*()
  **end if**
   Exploration thread
 **for** *workers* = 1 to *U* **do**
  Reset the initial state of IES dispatch environment
  Generate random scenario
  **for** dispatch time step *t* = 1 to *T* **do**
   Observe state $s_t$
   Select energy dispatch action $a_t$ by old actor $\theta^{\mu'}$
   Execute action $a_t$
   Calculate state of equipment by (13)-(38)
   Calculate reward $r_t$ by (41)
   Obtain the next state $s_{t+1}$
   *global_update_counter* += 1
   **if** *global_update_counter* > *mini_batch_size* **then**
    *exploration_event.clear*()
    *update_event.set*()
   **end if**
  **end for**
 **end for**
  Get trajectory $\tau$ and push data to chief
  **if not** *update_event.set*() **then**
   *update_event.wait*()
  **end if**
   Update thread
 **for** *m* = 1 to *M* **do**
  Calculate loss function $L(\theta^{Q})$ by (50)
  Update parameters of critic network $\theta^{Q}$ by (53)
  Calculate surrogate objective function $J(\theta^{\mu})$ by (46)
  Update parameters of new actor $\theta^{\mu}$ by (49)
  Update parameters of old actor: $\theta^{\mu'} \leftarrow \theta^{\mu}$
 **end for**
 *global_update_counter* = 0
 *update_event.clear*()
 *exploration_event.set*()
**end for**

TABLE III
PROPERTY PARAMETERS OF RPLT-CTM

| Parameter | Value | Parameter | Value |
|---|---|---|---|
| $\beta_e$ (t/MWh) | 1.08 | $\lambda_e$ (t/MWh) | 0.798 |
| $\beta_h$ (t/MWh) | 0.234 | $\lambda_h$ (t/MWh) | 0.385 |
| $\beta_{gas}$ (t/m$^3$) | $2.166\times10^{-3}$ | $\lambda_{gas}$ (t/m$^3$) | $1.95\times10^{-3}$ |
| $\beta_{PtG}$ (t/MWh) | 0.106 | | |

The parameters of the equipment operating constraints are provided in Table IV. The energy storage equipment parameters are provided in Table V.

TABLE IV
PARAMETERS OF EQUIPMENT OPERATING CONSTRAINTS

| Equipment | The minimum power (MW) | The maximum power (MW) | Climbing power (MW) |
|---|---|---|---|
| CHP | 0.2 | 1.2 | 0.1250 |
| PtG | 0.0 | 0.5 | 0.0625 |
| EB | 0.0 | 0.6 | 0.0750 |
| GB | 0.0 | 0.6 | 0.0750 |

TABLE V
ENERGY STORAGE EQUIPMENT PARAMETERS

| Equipment | Capacity (MWh) | Charging efficiency | Discharging efficiency |
|---|---|---|---|
| BES | 0.30 | 0.92 | 0.85 |
| TST | 0.30 | 0.95 | 0.95 |
| GST | 0.54 | 0.98 | 0.98 |

### B. Algorithm Setup

The proposed method and compared algorithms were implemented using TensorFlow and MATLAB. Simulation experiments were performed on a server with an Intel Xeon Gold 6230R CPU and an NVIDIA Quadro RTX 5000 GPU.

The core hyperparameter settings used for training the DPPO algorithm are listed in Table VI. The Adam optimizer is used to update the weights and biases of the actor and critic networks. The actor and critic networks contain two hidden layers with 300 and 100 neurons, respectively. All hidden layers use the rectified linear unit (ReLU) activation function.

TABLE VI
CORE HYPERPARAMETER SETTINGS FOR TRAINING DPPO ALGORITHM

| Hyperparameter | Value |
|---|---|
| Learning rate for actor network | 0.0001 |
| Learning rate for critic network | 0.0002 |
| Discount factor | 0.97 |
| The maximum episode | 10000 |
| Step in each episode | 96 |
| Mini-batch size | 64 |
| Surrogate objective function clipping rate | 0.2 |
| Number of parallel workers | 4 |

### C. Training Process

The DRL environment used to train the agent to learn a low-carbon economy dispatch policy was implemented based on Python 3.6, the framework of which is described in detail in Section III.

To verify the effectiveness of the established environment, an agent is trained in it using the DPPO algorithm. After testing different combinations of hyperparameters, the training results for the original version of the DRL environment are found to be poor. Therefore, to achieve better training results, state normalization (whitening) and reward normalization (whitening) are introduced. The cumulative rewards obtained from training in environments in which different tricks are applied are shown in Fig. 6.

In Fig. 6, the legend "None" represents the original version of the environment; "With state_norm" represents the environment with state normalization; "With reward_norm" represents the environment with reward normalization; and "With state_norm & reward_norm" represents the environment that uses both state normalization and reward normalization. The rewards obtained from the training show that the convergence of the algorithm cannot be improved de-

spite the use of reward normalization. The reward value remains low and fluctuates significantly. This result indicates that the agent does not learn an effective scheduling strategy. The reward value almost converges between 1500 and 3000 rounds. However, convergence is not maintained during the subsequent training process. When both state normalization and reward normalization are used, the reward value quickly converges and remains stable.

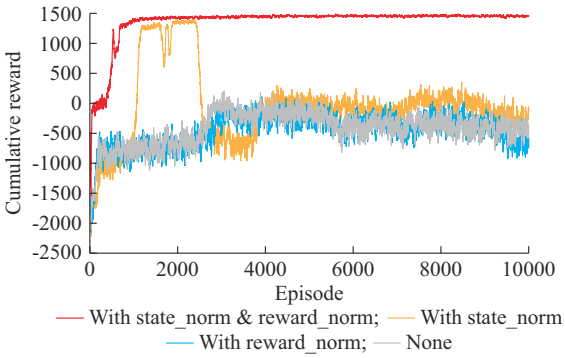


Fig. 6. Comparison of cumulative rewards in DRL environments with tricks.

By comparing and analyzing the training results of different environments, we notice that in the environment established in this study, the actor network and critic network are more suitable for the input-normalized states. In addition, the normalization of the reward helps the DRL agent to learn the dispatch strategy more effectively.

### *D. Simulation Result Analysis*

#### *1) Analysis on Results Based on Two Scenarios*

To analyze the benefits of introducing the RPLT-CTM for the low-carbon economic operation of IES, two scenarios are set up for comparative analysis, which are described as follows.

1) Scenario 1: the CTM is a carbon tax model in which the price of buying or selling carbon rights is fixed and does not change with the volume of carbon rights traded.

2) Scenario 2: the CTM is the RPLT-CTM model, the details of which are described in Section II.

To demonstrate the effectiveness of the proposed method, the actual operational data of an IES [37] are used for verification. The power load, heat load, gas load, and renewable energy generation power are presented in Fig. 7.

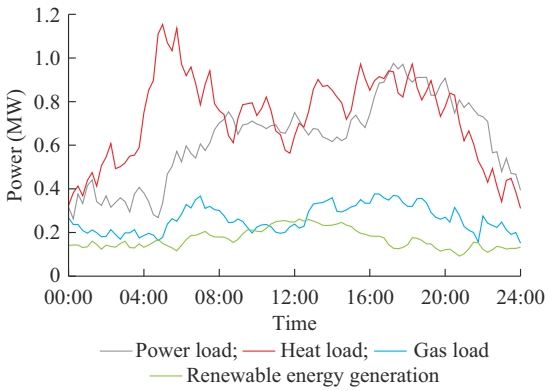


Fig. 7. Load demand and renewable energy generation on test day.

To intuitively compare the characteristics of the carbon trading models, the agent trained based on the DPPO algorithm in the two scenarios provides the scheduling plan according to the agent trained based on the DPPO algorithm in the two scenarios shown in Fig. 8. The scheduling results for the two scenarios, including the system operating costs and carbon emissions, are shown in Table VII.

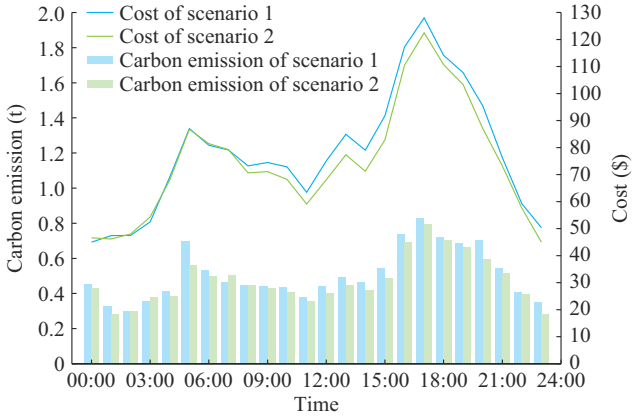


Fig. 8. Operating costs and carbon emission based on two scenarios.

TABLE VII
SCHEDULING RESULTS OF TWO SCENARIOS

| Scenario | Carbon credit (t) | Carbon emission (t) | Carbon trading cost ($) | Operating cost ($) |
|---|---|---|---|---|
| Scenario 1 | 15.89 | 12.16 | –179.22 | 1872.00 |
| Scenario 2 | 15.54 | 11.38 | –224.99 | 1789.24 |

Evidently, Fig. 8 clearly shows that the operating costs and carbon emissions of Scenario 1 are higher than those of Scenario 2. The reason for this result is that the carbon trading price of the carbon emission model in Scenario 1 is fixed and does not change with the accumulation of carbon trading volume. In Scenario 2, the carbon trading price changes in a stepwise manner with the accumulation of carbon trading volume, and the agent can develop a better scheduling plan under the guidance of such a mechanism. The carbon price gradually increases with the total amount of carbon rights purchased or sold. The purchase of carbon rights makes the system more expensive to operate, and the agent receives a penalty signal from the environment. The proceeds from the sale of carbon rights cut the system's operating costs, and the agent receives a reward signal from the environment. This price mechanism, which is punitive or rewarding in nature, can guide the agent in learning scheduling strategies that can reduce carbon emissions.

#### *2) Analysis on DPPO-based Method in Scenario 2*

The dispatch results of the IES based on DPPO for the test day in Scenario 2 are shown in Fig. 9. In Fig. 9(a), during the valley tariff period (00:00-08:00), the IES actively purchases power from the external grid and supplies the excess power to the PtG, EB, and BES systems. The PtG system converts electric power into natural gas power to supply the natural gas network, and the EB consumes electric power to provide heat power to the heating network. In addition, dur-

ing peak tariff periods (12:00-20:00), the IES also purchases power from the external grid to meet the demand for electrical loads of customers that cannot be met by equipment within the system, thereby ensuring a balance between the electric power supply and demand.

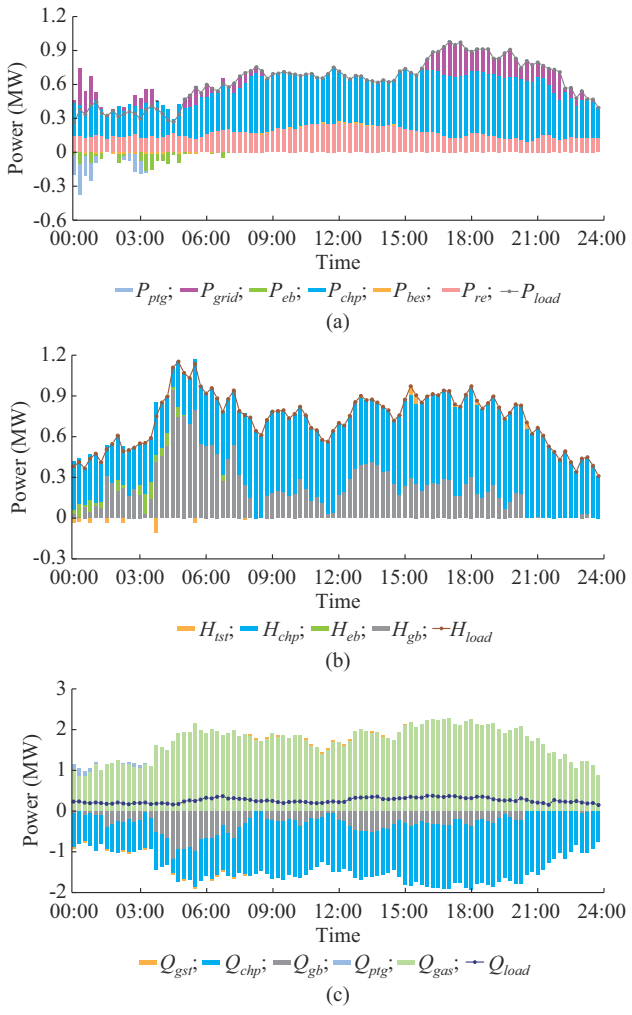


Fig. 9. Dispatch results of IES based on DPPO for test day in Scenario 2. (a) Electrical network. (b) Heating network. (c) Natural gas network.

In Fig. 9(b), to achieve economic operation of the system, the EB operates mainly during the valley tariff period (00:00-08:00). Although it is less expensive to run the EB during this period, the IES does not use the EB to provide a significant amount of thermal energy given the carbon emissions. During the period of 05:00-07:00, the heat load demand of customers is high, and to meet the load demand, the GB outputs a large amount of heat energy to supply to the heating network. The TST also outputs stored heat to the heating network when the heat load demand is high. Figure 9(c) shows the dispatch results of the natural gas network, where the CHP unit and GB consume large amounts of gas as load, and the PtG unit can supply natural gas to the network during the valley tariff to reduce operating costs.

Guided by the RPLT-CTM, the agent selects a dispatch plan with low carbon emissions and high economic efficiency. The detailed analysis of the scheduling results shows that the DPPO-trained dispatch agent provides real-time dispatch results according to the load demand and can achieve low-carbon and economic operation of the system by ensuring the safe and stable operation of the IES.

*3) Algorithm Comparison*

To verify the performance of the DPPO algorithm, DPPO algorithm is compared with other DRL algorithms and traditional algorithms in this subsection.

Since DPPO is a distributed version of PPO, PPO is chosen for comparison. The benchmark DRL algorithms, DDPG and twin-delayed DDPG (TD3), are selected. SAC, another popular DRL algorithm, is also used for comparison. Considering that DPPO is a distributed DRL algorithm, A3C and distributed distributional deterministic policy gradients (D4PG) are also introduced. In addition, the double deep $Q$-network (DDQN), an improved extension of the DQN algorithm, is employed as another benchmark DRL algorithm.

The cumulative rewards of DPPO and other DRL algorithms in the training process are shown in Fig. 10. DPPO converges quickly, reaching convergence after approximately 1200 episodes of training. In addition, DPPO obtains the highest cumulative reward among all selected DRL algorithms. D4PG and A3C, two distributed DRL algorithms, also converge quickly and reach convergence within 2000 episodes. TD3 also converges very quickly; however, it has a lower cumulative reward value than DPPO and D4PG. DDPG and PPO converge slowly, but receive higher rewards than A3C when they converge. The training results of SAC are poor, only better than those of DDQN in discrete action spaces. The comparison shows that DPPO is more efficient than the other DRL algorithms in learning to explore the optimal policy. In particular, the advantages of DPPO's distributed architecture are validated in comparison with PPO. Furthermore, DPPO obtains higher cumulative rewards in the convergence state, indicating that the algorithm learns to achieve a better strategy.

In addition, PSO-, GA-, and SO-based scheduling algorithms are introduced to compare IES operating costs and carbon emissions. The operating costs and carbon emissions of the scheduling plans for the test day provided by these algorithms are listed in Table VIII. Among them, the daily operating cost of the dispatch plan provided by DPPO is \$1789.24, which is 1.53%, 1.71%, 2.13%, 2.84%, 3.76%, 10.99%, 12.40%, 4.84%, 5.28%, and 3.82% lower than that of D4PG, TD3, PPO, DDPG, A3C, SAC, DDQN, GA, PSO, and SO, respectively. The daily carbon emission of the dispatch plan given by the DPPO-based method is 11.38 t, which is 5.09%, 6.49%, 3.23%, 6.87%, 14.95%, 21.08%, 28.16%, 21.03%, 15.83%, and 8.81% lower than that of D4PG, TD3, PPO, DDPG, A3C, SAC, DDQN, GA, PSO, and SO, respectively.

The results show that DRL-based dispatch algorithms with a continuous action space outperform the PSO- and SO-based algorithms. This is a consequence of the fact that DRL-based dispatch algorithms do not rely on day-ahead forecast information or an assumed distribution of uncertainty. In contrast, the DRL-based algorithm (DDQN) with a discrete action space is limited to a finite number of actions available in the action space. Therefore, its scheduling results are the worst among all algorithms.

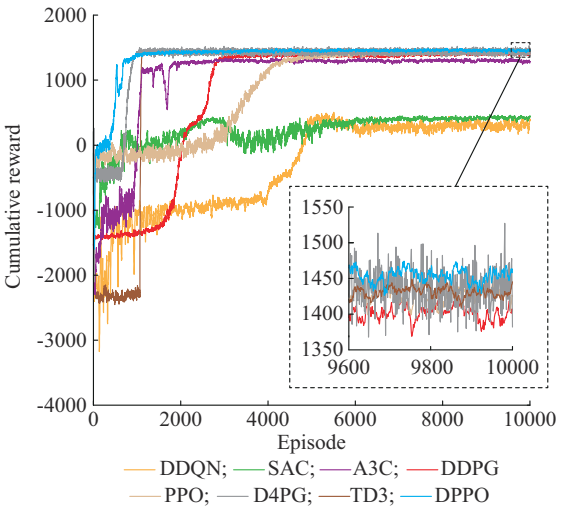


Fig. 10. Cumulative rewards of DPPO and other DRL algorithms in training process.

TABLE VIII
SCHEDULING RESULTS USING DIFFERENT ALGORITHMS

| Algorithm | Carbon credit (t) | Carbon emission (t) | Carbon trading cost ($) | Operating cost ($) |
|---|---|---|---|---|
| DPPO | 15.54 | 11.38 | –224.99 | 1789.24 |
| D4PG | 15.94 | 11.99 | –220.42 | 1817.08 |
| TD3 | 15.99 | 12.17 | –214.79 | 1820.42 |
| PPO | 15.95 | 11.76 | –219.44 | 1828.25 |
| DDPG | 16.34 | 12.22 | –217.19 | 1841.59 |
| A3C | 16.95 | 13.38 | –188.78 | 1859.15 |
| SAC | 17.19 | 14.42 | –152.67 | 2010.06 |
| DDQN | 18.07 | 15.84 | –133.29 | 2042.46 |
| GA | 17.90 | 14.41 | –191.01 | 1880.23 |
| PSO | 17.74 | 13.52 | –222.37 | 1889.07 |
| SO | 16.79 | 12.48 | –224.83 | 1860.24 |

The above analysis suggests that the DPPO-based method has higher learning efficiency and a better dispatch strategy than the other DRL-based algorithms. A comparison with other dispatch algorithms shows that the DPPO-based method also provides a better dispatch strategy.

## V. CONCLUSION

In this paper, considering the uncertainty of load demand and renewable energy, a low-carbon ED method for electricity-heat-gas IES based on DRL is proposed. A reward function based on the RPLT-CTM is introduced to guide the DRL agent to learn low-carbon dispatch actions. A DRL agent trained by DPPO realizes the real-time low-carbon ED of an IES. The following conclusions are drawn.

1) Benefiting from the ladder-type dynamic trading price, the RPLT-CTM effectively guides the DRL agent to learn a low-carbon ED strategy. The dispatch results verify that the agent based on the RPLT-CTM makes a dispatch plan with lower carbon emissions compared with the agent based on the traditional CTM.

2) The effectiveness of the proposed DRL-based method for low-carbon ED of an electricity-heat-gas IES is demonstrated by the dispatch results on the test day. The agent trained using the proposed method controls the dispatch actions of each device in the IES in real time. The dispatch plan generated by the agent achieves the low-carbon economic operation of the electricity-heat-gas IES.

3) The superiority of DPPO is verified through a comparative analysis. The distributed architecture of DPPO enables it to perform better than PPO in complex training environments. Compared with the scheduling results of PPO, DPPO reduces the operating cost and carbon emissions by 2.13% and 3.23%, respectively. Compared with other distributed DRL algorithms (D4PG and A3C), the operating cost and carbon emissions of the DPPO-based method are reduced by 1.53%, 3.76% and 5.09%, 14.95%, respectively. DPPO is also compared with other DRL algorithms (DDPG, A3C, SAC, and DDQN) and dispatch algorithms (GA, PSO, and SO). The operating costs of the DPPO-based dispatch method are reduced by 2.84%, 3.76%, 10.99%, 12.40%, 4.84%, 5.28%, and 3.82%, and the carbon emissions are reduced by 6.87%, 14.95%, 21.08%, 28.16%, 21.03%, 15.83%, and 8.81%, respectively.

In future work, considering the characteristics of multiple operators of IES, multi-agent reinforcement learning will be applied to the optimal operation of an IES.

## REFERENCES

[1] P. Li, Z. Wang, J. Wang *et al.*, "Two-stage optimal operation of integrated energy system considering multiple uncertainties and integrated demand response," *Energy*, vol. 225, p. 120256, Jun. 2021.
[2] Y. Li, M. Han, Z. Yang *et al.*, "Coordinating flexible demand response and renewable uncertainties for scheduling of community integrated energy systems with an electric vehicle charging station: a bi-level approach," *IEEE Transactions on Sustainable Energy*, vol. 12, no. 4, pp. 2321-2331, Oct. 2021.
[3] L. Chen, Q. Xu, Y. Yang *et al.*, "Community integrated energy system trading: a comprehensive review," *Journal of Modern Power Systems and Clean Energy*, vol. 10, no. 6, pp. 1445-1458, Nov. 2022.
[4] W. Wang, S. Huang, G. Zhang *et al.*, "Optimal operation of an integrated electricity-heat energy system considering flexible resources dispatch for renewable integration," *Journal of Modern Power Systems and Clean Energy*, vol. 9, no. 4, pp. 669-710, Jul. 2021.
[5] W. Wang, S. Huang, G. Zhang *et al.*, "Optimal operation of an integrated electricity-heat energy system considering flexible resources dispatch for renewable integration," *Journal of Modern Power Systems and Clean Energy*, vol. 9, no. 4, pp. 699-710, Jul. 2021.
[6] R. Rocchetta, L. Bellani, M. Compare *et al.*, "A reinforcement learning framework for optimal operation and maintenance of power grids," *Applied Energy*, vol. 241, pp. 291-301, May 2019.
[7] A. T. D. Perera and P. Kamalaruban, "Applications of reinforcement learning in energy systems," *Renewable and Sustainable Energy Reviews*, vol. 137, p. 110618, Mar. 2021.
[8] T. Yang, L. Zhao, W. Li *et al.*, "Reinforcement learning in sustainable energy and electric systems: a survey," *Annual Reviews in Control*, vol. 49, pp. 145-163, Apr. 2020.
[9] L. He, Z. Lu, J. Zhang *et al.*, "Low-carbon economic dispatch for electricity and natural gas systems considering carbon capture systems and power-to-gas," *Applied energy*, vol. 224, pp. 357-370, Aug. 2018.
[10] H. Vella, "Last chance for carbon trading? Leaders at the COP26 climate conference will consider how to create a framework for global cooperation on carbon markets, which could be a key breakthrough for climate change mitigation," *Engineering & Technology*, vol. 16, no. 10, pp. 1-4, Nov. 2021.
[11] The People's Government of Hainan Province. (2023, Jan.). Hainan International Carbon Emission Trading Center achieved its first cross-border carbon trading. [Online]. Available: https://www.hainan.gov.cn/hainan/5309/202301/7a3d3c12136f43e986b95578dd90de08.shtml
[12] Y. Li, Y. Zou, Y. Tan *et al.*, "Optimal stochastic operation of integrat-

ed low-carbon electric power, natural gas, and heat delivery system," *IEEE Transactions on Sustainable Energy*, vol. 9, no. 1, pp. 273-283, Jan. 2018.
[13] S. Lu, W. Gu, S. Zhou *et al.*, "Adaptive robust dispatch of integrated energy system considering uncertainties of electricity and outdoor temperature," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 7, pp. 4691-4702, Jul. 2020.
[14] A. Mansour-Saatloo, Y. Pezhmani, M. A. Mirzaei *et al.*, "Robust decentralized optimization of multi-microgrids integrated with power-to-X technologies," *Applied Energy*, vol. 304, p. 117635, Dec. 2021.
[15] N. Nasiri, S. Zeynali, S. N. Ravadanegh *et al.*, "A hybrid robust-stochastic approach for strategic scheduling of a multi-energy system as a price-maker player in day-ahead wholesale market," *Energy*, vol. 235, p. 121398, Nov. 2021.
[16] M. A. Mirzaei, K. Zare, B. Mohammadi-Ivatloo *et al.*, "Robust network-constrained energy management of a multiple energy distribution company in the presence of multi-energy conversion and storage technologies," *Sustainable Cities and Society*, vol. 74, p. 103147, Nov. 2021.
[17] Y. Zhang, F. Zheng, S. Shu *et al.*, "Distributionally robust optimization scheduling of electricity and natural gas integrated energy system considering confidence bands for probability density functions," *International Journal of Electrical Power & Energy Systems*, vol. 123, p. 106321, Dec. 2020.
[18] X. Lu, Z. Liu, L. Ma *et al.*, "A robust optimization approach for optimal load dispatch of community energy hub," *Applied Energy*, vol. 259, p. 114195, Feb. 2020.
[19] Z. Li, L. Wu, Y. Xu *et al.*, "Multi-stage real-time operation of a multi-energy microgrid with electrical and thermal energy storage sets: a data-driven MPC-ADP approach," *IEEE Transactions on Smart Grid*, vol. 13, no. 1, pp. 213-226, Jan. 2022.
[20] X. Jin, Q. Wu, H. Jia *et al.*, "Optimal integration of building heating loads in integrated heating/electricity community energy systems: a bi-level MPC approach," *IEEE Transactions on Sustainable Energy*, vol. 12, no. 3, pp. 1741-1754, Jul. 2021.
[21] N. Nasiri, S Zeynali, S. N. Ravadanegh *et al.*, "A tactical scheduling framework for wind farm-integrated multi-energy systems to take part in natural gas and wholesale electricity markets as a price setter," *IET Generation, Transmission & Distribution*, vol. 16, no. 9, pp. 1849-1864, Feb. 2022.
[22] A. Mansour-Saatloo, R. Ebadi, M. A. Mirzaei *et al.*, "Multi-objective IGDT-based scheduling of low-carbon multi-energy microgrids integrated with hydrogen refueling stations and electric vehicle parking lots," *Sustainable Cities and Society*, vol. 74, p. 103197, Nov. 2021.
[23] Y. Ji, J. Wang, J. Xu *et al.*, "Real-time energy management of a microgrid using deep reinforcement learning," *Energies*, vol. 12, no. 12, p. 2291, Jun. 2019.
[24] Y. Liu, D. Zhang, and H. B. Gooi, "Optimization strategy based on deep reinforcement learning for home energy management," *CSEE Journal of Power and Energy Systems*, vol. 6, no. 3, pp. 572-582, Sept. 2020.
[25] F. Meng, Y. Bai, and J. Jin, "An advanced real-time dispatching strategy for a distributed energy system based on the reinforcement learning algorithm," *Renewable Energy*, vol. 178, pp. 13-24, Nov. 2021.
[26] K. Zhou, K. Zhou, and S. Yang, "Reinforcement learning-based scheduling strategy for energy storage in microgrid," *Journal of Energy Storage*, vol. 51, p. 104379, Jul. 2022.
[27] E. Mocanu, D. C. Mocanu, P. H. Nguyen *et al.*, "On-line building energy optimization using deep reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 10, no. 4, pp. 3698-3708, May 2018.
[28] T. A. Nakabi and P. Toivanen, "Deep reinforcement learning for energy management in a microgrid with flexible demand," S*ustainable Energy, Grids and Networks*, vol. 25, p. 100413, Mar. 2021.
[29] L. Lei, Y. Tan, G. Dahlenburg *et al.*, "Dynamic energy dispatch based on deep reinforcement learning in IoT-driven smart isolated microgrids," *IEEE Internet of Things Journal*, vol. 8, no. 10, pp. 7938-7953, May 2021.
[30] C. Guo, X. Wang, Y. Zheng *et al.*, "Real-time optimal energy management of microgrid with uncertainties based on deep reinforcement learning," *Energy*, vol. 238, p. 121873, Jan. 2022.
[31] B. Zhang, W. Hu, D. Cao *et al.*, "Deep reinforcement learning-based approach for optimizing energy conversion in integrated electrical and heating system with renewable energy," *Energy Conversion and Management*, vol. 202, p. 112199, Dec. 2019.
[32] S. Zhou, Z. Hu, W. Gu *et al.*, "Combined heat and power system intelligent economic dispatch: a deep reinforcement learning approach," *International Journal of Electrical Power & Energy Systems*, vol. 120, p. 106016. Sept. 2020.
[33] T. Yang, L. Zhao, W. Li *et al.*, "Dynamic energy dispatch strategy for integrated energy system based on improved deep reinforcement learning," *Energy*, vol. 235, p. 121377, Nov. 2021.
[34] Y. Ye, D. Qiu, X. Wu *et al.*, "Model-free real-time autonomous control for a residential multi-energy system using deep reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 11, no. 4, pp. 3068-3082, Jul. 2020.
[35] L. Zhao, T. Yang, W. Li *et al.*, "Deep reinforcement learning-based joint load scheduling for household multi-energy system," *Applied Energy*, vol. 324, p. 119346, Oct. 2022.
[36] B. Zhang, W. Hu, J. Li *et al.*, "Dynamic energy conversion and management strategy for an integrated electricity and natural gas system with renewable energy: deep reinforcement learning approach," *Energy Conversion and Management*, vol. 220, p. 113063, Sept. 2020.
[37] J. Dong, H. Wang, J. Yang *et al.*, "Optimal scheduling framework of electricity-gas-heat integrated energy system based on asynchronous advantage actor-critic algorithm," *IEEE Access*, vol. 9, pp. 139685-139696, Sept. 2021.
[38] Q. Sun, D. Wang, D. Ma *et al.*, "Multi-objective energy management for we-energy in Energy Internet using reinforcement learning," in *Proceedings of 2017 IEEE Symposium Series on Computational Intelligence*, Honolulu, USA, Dec. 2017, pp. 1-6.
[39] X. Teng, H. Long, and L. Yang, "Integrated electricity-gas system optimal dispatch based on deep reinforcement learning," in *Proceedings of IEEE Sustainable Power and Energy Conference*, Nanjing, China, Dec. 2021, pp. 1082-1086.
[40] B. Zhang, W. Hu, D. Cao *et al.*, "Soft actor-critic-based multi-objective optimized energy conversion and management strategy for integrated energy systems with renewable energy," *Energy Conversion and Management*, vol. 243, p. 114381, Sept. 2021.
[41] G. Zhang, W. Hu, D. Cao *et al.*, "A multi-agent deep reinforcement learning approach enabled distributed energy management schedule for the coordinate control of multi-energy hub with gas, electricity, and freshwater," *Energy Conversion and Management*, vol. 255, p. 115340, Mar. 2022.
[42] T. Chen, S. Bu, X. Liu *et al.*, "Peer-to-peer energy trading and energy conversion in interconnected multi-energy microgrids using multi-agent deep reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 13, no. 1, pp. 715-727, Jan. 2022.
[43] D. Qiu, Z. Dong, X. Zhang *et al.*, "Safe reinforcement learning for real-time automatic control in a smart energy-hub," *Applied Energy*, vol. 309, p. 118403, Mar. 2022.
[44] Q. Sun, X. Wang, Z. Liu *et al.*, "Multi-agent energy management optimization for integrated energy systems under the energy and carbon co-trading market," *Applied Energy*, vol. 324, p. 119646, Oct. 2022.
[45] D. Qiu, J. Xue, T. Zhang *et al.*, "Federated reinforcement learning for smart building joint peer-to-peer energy and carbon allowance trading," *Applied Energy*, vol. 333, p. 120526, Mar. 2023.
[46] R. Wang, X. Wen, X. Wang *et al.*, "Low carbon optimal operation of integrated energy system based on carbon capture technology, LCA carbon emissions and ladder-type carbon trading," *Applied Energy*, vol. 311, p. 118664, Apr. 2022.
[47] X. Zhang, X. Liu, J. Zhong *et al.*, "Electricity-gas-integrated energy planning based on reward and penalty ladder-type carbon trading cost," *IET Generation, Transmission & Distribution*, vol. 13, no. 23, pp. 5263-5270, Dec. 2019.
[48] J. Schulman, F. Wolski, P. Dhariwal *et al.* (2017, Jul.). Proximal policy optimization algorithms. [Online]. Available: https://arxiv. org/abs/1707.06347
[49] N. Heess, T. B. Dhruva, S. Sriram *et al.* (2017, Jul.). Emergence of locomotion behaviours in rich environments. [Online]. Available: https://arxiv.org/abs/1707.02286

**Yuxian Zhang** received the M.S. and Ph.D. degrees in control theory and control engineering at Northeastern University, Shenyang, China, in 2005 and 2007, respectively. He worked in postdoctoral station of control science and engineering at Tsinghua University, Beijing, China, in 2007-2009. He joined Shenyang University of Technology, Shenyang, China, in 2009. His current research interests include integrated energy system, power system operation analysis and dispatching, and load modeling.

**Yi Han** received the B.E. degree in new energy science and engineering at Shenyang Institute of Engineering, Shenyang, China, in 2019. He is currently pursuing the M.S. degree in electrical engineering in Shenyang University of Technology, Shenyang, China. His research interests include integrated

energy system, power system operation analysis and dispatching.

**Deyang Liu** received the B.E. degree in automation and the M.S. degree in electrical engineering from Shenyang University of Technology, Shenyang, China, in 2010 and 2013, respectively. He is currently pursuing the Ph.D. degree in electrical engineering in Shenyang University of Technology. His research interests include integrated energy system and fault diagnosis.

**Xiao Dong** received the B.E. degree in automation and the M.S. degree in electrical engineering from Shenyang University of Technology, Shenyang, China, in 2009 and 2012, respectively. She works in Beijing Ke Dong Co., Ltd., NARI Group Corporation, Beijing, China. Her research interests include power system operation analysis and dispatching.