Optimal Joint Operation Method of Integrated Electricity and Heating Systems Based on Multi-agent Deep Reinforcement Learning Method

Hangyue Liu, Cuo Zhang, Jiawei Wang, Ke Meng, and Zhao Yang Dong

***Abstract*—Conventional joint operation of integrated electricity and heating systems faces severe challenges, including non-convex models and computation complexity. Additionally, there are adverse impacts from the uncertainties of renewable distributed generators, as well as electrical and thermal loads. This paper proposes an optimal joint operation method of integrated electricity and heating systems based on multi-agent deep reinforcement learning (DRL) method. Firstly, a new hydraulic-thermal flow algorithm that is compatible with DRL training environment is developed. Then, a stochastic distributed optimization model is formulated with multiple agents to minimize network power losses while avoiding operation constraint violations under the spatial and temporal uncertainties. Last, a multi-agent deep deterministic policy gradient is adopted combined with offline neural network training via exploration in a virtual environment and online optimization of joint operation. A numerical case study indicates the effectiveness of the proposed method and solution robustness against spatial and temporal uncertainties.**

***Index Terms*—Deep reinforcement learning (DRL), hydraulic-thermal flow, power distribution system, neural network, agont, integrated electricity and heating system, uncertainey.**

I. Introduction

With increasing renewable distributed generators (RDGs) and the global transition towards low carbon energy, integrated electricity and heating systems (IEHSs) have been widely adopted [1]. The conventional isolated energy systems only consider the optimization of operation within each system and cannot be complementarily coordinated to boost energy efficiency [2]. Thus, in an IEHS, power distribution systems (PDSs) and district heating systems (DHSs) are expected to be optimally coordinated, achieving an optimal joint operation with high energy utilization efficiency.

It is imperative to optimally coordinate different energy systems with flexible control on coupling units such as combined heat and power (CHP) plant that can simultaneously supply thermal and electrical energy. With the control on CHP plants and assessing respective energy flows, PDS and DHS can be jointly operated. Compared with conventional power flow only focusing on power system operation, the joint operation of PDS and DHS requires the modeling of CHP flow in the power system and DHS. The non-convex thermal energy model and the coordination of two energy flows bring extra challenge to the IEHS operation. In literature, the operation of CHP plants in a distribution-level IEHS is investigated in [3], where a sequential solution algorithm for a hydraulic-thermal power flow model is developed. Besides, the dynamic process of heat flow and the issue of time delay are considered in [4], where the partial differential equation of the temperature is modeled.

However, these studies focus on the modeling of multiple energy flows, while various uncertainties impacting the operation are ignored. Uncertainties arise from RDGs as well as thermal and electrical loads, and they can significantly impact joint operation decision-making in terms of economic objectives and technical constraints. Reference [5] considers uncertainties of renewable and multi-energy loads in multi-energy microgrid operation and addresses them via a robust optimization method, but the study focuses on aggregated thermal loads rather than DHSs. Reference [6] proposes a two-stage cost-effective operation method for a CHP plant and wind farm generation portfolio considering the uncertainties of thermal loads and wind power generation via stochastic programming. However, the stochastic optimization method in multiple scenarios is computationally expensive. For a

Manuscript received: March 30, 2025; revised: June 7, 2025; accepted: August 4, 2025. Date of CrossCheck: August 4, 2025. Date of online publication: November 25, 2025.

This work was partially supported by JC STEM Lab of Future Energy Systems (No. 2025-0039), Global STEM Professorship (No. GSP313), China Postdoctoral Science Foundation (No. 2024M750373), and National Natural Science Foundation of China (No. 62503103).

H. Liu is with Sungrow Power Supply Co., Ltd., Hefei, China (e-mail: henry-liu0616@gmail.com).

C. Zhang is with School of Electrical and Computer Engineering, The University of Sydney, Sydney, Australia (e-mail: cuo.zhang@sydney.edu.au).

J. Wang (corresponding author) is with National Frontiers Science Center for Industrial Intelligence and Systems Optimization, Northeastern University, Shenyang, China (e-mail: wangjiawei@mail.neu.edu.cn).

K. Meng is with School of Electrical Engineering and Telecommunications, The University of New South Wales, Sydney, Australia (e-mail: ke.meng@unsw.edu.au).

Z. Y. Dong is with Department of Electrical Engineering, City University of Hong Kong, Hong Kong, China (e-mail: zydong@ieee.org).

DOI: 10.35833/MPCE.2025.000285

PDS, besides spatial uncertainties, temporal uncertainties due to intermittent power generation and consumption as well as communication/response delays should be modeled and addressed.

The joint operation optimization itself has computation challenges due to high nonlinearity of both hydraulic-thermal flow and power flow. In a DHS, the computation difficulty comes from bilinear terms of water temperature and mass flow rate. Conventional research works including [7] apply the Newton-Raphson algorithm to solve the hydraulic-thermal flow iteratively with heavy computation burdens. The iteration method is simplified with a fixed supply temperature of CHP plants to reduce the computation burdens, while the supply temperature should be controlled by the plant operator for economic system operation. Recently, a commercial nonlinear solver is used in [8]. In [9], a piecewise linearization technique is utilized for the non-linear constraints. In addition, the convex relaxation methods such as a convex-concave procedure method [10] are utilized to deal with the bilinear term. However, the optimality gap may not be guaranteed, and the computation complexity is still very high. For a PDS, an accurate AC power flow model also requires the iterative Newton-Raphson algorithm. With complex operation conditions under heavy uncertainties, power system optimization models with relaxations may have potential side effects of reduced accuracy.

To address these challenges, deep reinforcement learning (DRL) methods provide potential. As a data-driven online optimization method, DRL methods can learn the optimal decision policies while agents continuously interact with the environment where multiple uncertainties are applied, without the prior knowledge of the probability distributions of those uncertainties [11]. Deep learning plays a crucial role in DRL methods by learning the mapping from states to actions. Deep learning models have been applied to solve the optimal power flow in power system operation [12] and volt/var control problem [13]. The DRL methods have been applied for power system operation [14]. Results indicate that compared with conventional stochastic programming methods, the DRL methods can contribute to reducing the operation costs and satisfying operation constraints in a power system [14]. While the mentioned challenges are mitigated in these research works, these DRL methods are applied for centralized operation frameworks suffering from scalability and privacy issues. To this point, multi-agent DRL method has been developed to solve operation optimization problems of large-scale systems [15].

DRL methods, especially those with off-policy algorithms that can provide optimal policy, are efficient in solving optimization problems [16]. With stable convergence performance as a major advantage, deep deterministic policy gradient (DDPG) algorithm has been widely used, and [17] develops a multi-agent DDPG (MADDPG) optimization model for volt/var control of active distribution networks under various uncertainties. The DDPG is suggested for its high sampling efficiency against on-policy algorithms and fewer parameters used during training [18]. The DDPG is also able to handle continuous action spaces in an environment such as power system operation, IEHS energy management [19] [20], and demand response [21]. Thus, DDPG optimization method can be developed for IEHS operation.

An actor-critic-based DRL method is proposed in [22] to solve an energy management problem of a smart home. However, the models of electrical power and heating networks are not considered. Such network modeling is critical for system-level IEHS operation, where the operation characteristics from both networks are coupled and should be monitored by system operators for secure operation. A proximal policy optimization (PPO) method is developed in [23] to minimize the operation cost of a medium system-level IEHS, with the IEEE 15-bus test feeder. The method demonstrates reliable performance against uncertainties in comparison with conventional mathematical methods including stochastic programming. In [24], compared with the PPO method, the DDPG algorithm outperforms in keeping operation constraints such as system energy balancing, and it is less sensitive to various hyperparameters. Reference [25] adopts a DDPG algorithm to minimize the daily operation cost of an IEHS. However, the system operation conditions are simplified via a power balance model, so that the power losses and voltage dynamics are still omitted. On the other hand, [22] - [25] rely on single-agent DRL method, and the IEHS size is limited to one single heat and power source.

For a large-scale IEHS including multiple sub-systems, the single-agent DRL method can lead to high dimensional action and state space. The training process can suffer from slow convergence and high computation complexity. With a multi-agent DRL method, each local system operator can be responsible for its own network operation without sharing operation information with each other [26]. Considering the advantages of DDPG algorithm mentioned above, the MADDPG algorithm has been investigated to address these issues. Reference [27] applies an MADDPG algorithm to the scheduling of microgrid-level IEHSs, which can optimize the energy costs and carbon emissions. However, in the above research works, DHS hydraulic-thermal flow and PDS power flow models are either not considered or simplified with fixed supply temperature, thus losing accuracy. It is worth noting that the existing hydraulic-thermal flow algorithms are not compatible with the DRL methods. Besides, the uncertainties caused by geographically distributed power generation, electrical and thermal loads, as well as temporal intermittencies have not been fully considered or specifically addressed.

To address the above unsolved issues, we propose an optimal joint operation method of IEHSs based on multi-agent DRL method, minimizing power and thermal losses for IEHSs with accurate hydraulic-thermal and power flow models. Multiple agents are used for distribution sub-networks and DHSs to achieve distributed operation, while various spatial and temporal uncertainties are addressed. The main contributions of the paper are given as follows.

1) A distributed optimization model is addressed for optimal joint operation of IEHSs, considering spatial and temporal uncertainties.

2) A new hydraulic-thermal flow algorithm is proposed

for DHS operation, where supply and return temperatures of CHP plants and mass flow rates are all optimized, which is compatible with DRL method.

3) An MADDPG method is applied, where each power distribution and district heating area with a CHP plant is optimized by an agent, to achieve a data-driven optimal joint operation of IEHSs.

II. Data-driven Optimal Joint Operation

This paper proposes a data-driven optimal joint operation method for IEHSs. Modern PDS operation model can minimize network power losses and/or voltage deviations by controlling inverter-based RDG. DHSs aim to minimize operation cost while fulfilling thermal requirement of the user, which is equivalent to the minimization of thermal losses while keeping water temperature within a range. Thus, the optimal joint operation of IEHSs aims to minimize power and thermal losses by controlling the CHP plants and RDGs.

To alleviate computation burdens and mitigate privacy issues, this paper proposes a distributed optimization framework with multiple agents, which is compatible with efficient multi-agent DRL method. The schematic system diagram of an IEHS with multiple agents is shown in Fig. 1.

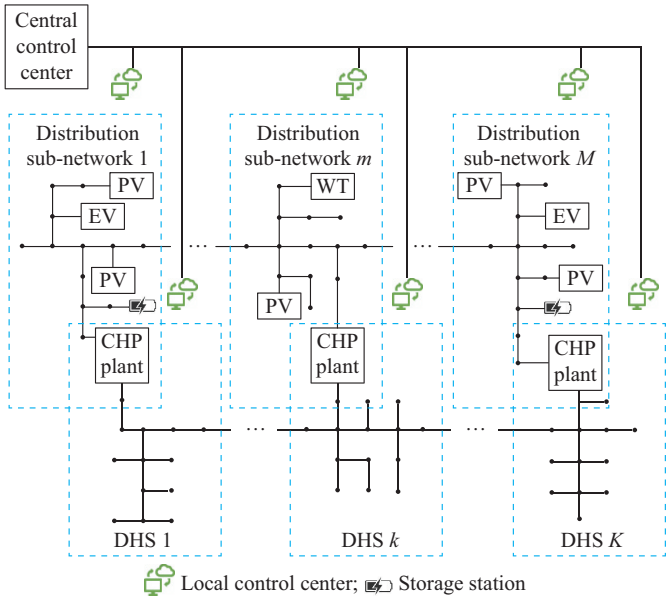

Fig. 1. Schematic system diagram of an IEHS with multiple agents.

The IEHS consists of a PDS and several DHSs. The PDS can be partitioned into multiple distribution sub-networks with local control centers. For each distribution sub-network, it is coupled with a DHS via a CHP plant. Each DHS has a local control center to adjust the electrical and thermal power outputs. Under this framework, the distribution sub-networks remain interconnected, whereas DHSs are isolated with a CHP plant as the thermal supply source. The hydraulic-thermal flow algorithm works for each individual DHS, while the power flow algorithm works for the entire PDS.

The PDS has a central control center with communication links to the local control centers. Applying the MADDPG method, this joint operation is characterized by centralized training and decentralized execution. The central control center takes the responsibility of neural network training for the whole IEHS, and it sends the well-trained neural network for each local control center. In real time, using the neural network, each local control center determines the operation actions with corresponding local measurements and dispatches them to the CHP plant or the RDGs. Note that each local control center is an agent for the MADDPG method.

III. Hydraulic-thermal Flow Algorithm

The DHS operation aims to provide sufficient thermal energy for district heating customers and fulfill the supply temperature requirements. For this purpose, this paper develops a new hydraulic-thermal flow algorithm, which takes thermal supply of CHP plants and thermal demand of customers as inputs and determines supply temperatures in a DHS. Then, the temperatures are checked if they are within corresponding required ranges. This algorithm is compatible with the DRL training environment.

A. Hydraulic-thermal Model

For DHS k, a hydraulic-thermal model can be formulated as:

$$\boldsymbol{A}_k \dot{\boldsymbol{m}}_{s,k} = \dot{\boldsymbol{m}}_{q,k} \tag{1}$$

$$\boldsymbol{B}_k \boldsymbol{E}_k \dot{\boldsymbol{m}}_{s,k} \left| \dot{\boldsymbol{m}}_{s,k} \right| = 0 \tag{2}$$

$$\left(\sum_{\forall n \in L_n^+} \dot{m}_n \right) T_n = \sum_{\forall n \in L_n^-} \dot{m}_n T_n \tag{3}$$

$$T_l^- = \left(T_l^+ - T_a \right) \mathrm{e}^{-\frac{\lambda_l \Lambda_l}{c_w \dot{m}_l}} + T_a \quad \forall l \tag{4}$$

$$\phi_n^d = c_w \dot{m}_{q,n} \left(T_n^{sup} - T_n^{rtn} \right) \tag{5}$$

$$\phi_{k,t}^{CHP} = c_w \dot{m}_{q,k} \left(T_k^{src} - T_k^{rtn} \right) \tag{6}$$

where subscript t is the time index; $\boldsymbol{A}_k$ is the network incidence matrix in DHS k; $\dot{\boldsymbol{m}}_{s,k}$ is the vector of the nodal mass flow rates in DHS k; $\dot{\boldsymbol{m}}_{q,k}$ is the vector of the nodal net injection mass flow rates in DHS k; $\boldsymbol{B}_k$ is the loop incidence matrix in DHS k; c_w is the specific heat of water; $\boldsymbol{E}_k$ is the resistance coefficient matrix of pipes in DHS k; $\dot{m}_{q,n}$ and $\dot{m}_{q,k}$ are the nodal mass flow rates of node n and DHS k, respectively; $\dot{m}_l$ and $\dot{m}_n$ are the mass flow rates of pipe l and node n, respectively; T_l^+ and T_l^- are the temperatures of start and end of pipe l, respectively; T_a is the ambient temperature; T_n is the temperature of node n; T_k^{src} is the temperature of source node in DHS k; T_k^{rtn} is the return temperature in DHS k; Λ_l is the friction factor of pipe l; λ_l is the length of pipe l; $\mathcal{L}_n^+$ and $\mathcal{L}_n^-$ are the water flows to pipe and from pipe, respectively; $\phi_{k,t}^{CHP}$ is the thermal power generation of CHP in DHS k; ϕ_n^d is the thermal power consumption; and T_n^{sup} and T_n^{rtn} are the supply and return temperatures at node n, respectively.

Equations (1) and (2) present the hydraulic mass flow rates of the DHS. The thermal energy equality is given by

(3), while the temperature drop of a pipe along water flow is expressed by (4). The thermal power and nodal mass flow rate for district heating demand and CHP plants are given by (5) and (6), respectively. Detailed explanation can be found in [28].

B. Hydraulic-thermal Flow Algorithm

A new hydraulic-thermal flow algorithm is developed and shown in Fig. 2.

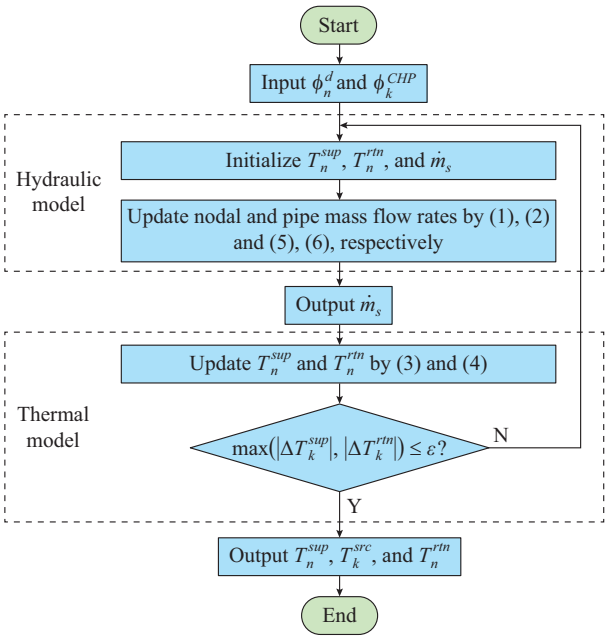

Fig. 2. Hydraulic-thermal flow algorithm.

This algorithm takes ϕ_n^d and ϕ_k^{CHP} as inputs. After the initialization of T_n^{sup}, T_n^{rtn}, and $\dot{m}_s$, the algorithm updates the nodal and pipe mass flow rates by (1) and (2) as well as (5) and (6), respectively. Then, T_n^{sup} and T_n^{rtn} are updated by (3) and (4). If the changes of supply and return temperatures ΔT_k^{sup} and ΔT_k^{rtn} are within the iteration termination threshold ε, the outputs, i.e., T_k^{src}, T_n^{sup}, and T_n^{rtn}, are obtained.

For the secure DHS operation, T_k^{src} should be limited by (7), while T_n^{sup} should also be controlled within the customer required range as shown in (8).

$$T_k^{src,\min} \le T_k^{src} \le T_k^{src,\max} \tag{7}$$

$$T_n^{sup,\min} \le T_n^{sup} \le T_n^{sup,\max} \quad \forall n \tag{8}$$

where $T_k^{src,\max}$ and $T_k^{src,\min}$ are the maximum and minimum source temperature limits, respectively; and $T_n^{sup,\max}$ and $T_n^{sup,\min}$ are the maximum and minimum supply temperature limits, respectively.

After obtaining T_k^{src} and T_n^{sup} by the hydraulic-thermal flow algorithm, the results are checked if they are within the required operation ranges. If the temperature results are out of the ranges, the CHP operator should modify the power output of CHP plant, thus providing sufficient thermal energy in the DHS. The hydraulic-thermal flow algorithm may not cover certain thermal power inputs, leading to vast mismatches between thermal supply and demand. Under this condition, it is expected to modify the thermal power output of the CHP plant first.

The developed hydraulic-thermal flow algorithm is compatible with the training process of DRL method. For training neural networks in a DRL method, random power outputs of CHP plant are generated and used as inputs in the hydraulic-thermal flow algorithm. Then, the network source and supply temperatures of each node are checked to adjust the penalty in reward functions of neural network training, such that the temperature constraint violations can be avoided after training.

IV. Optimization Model Under Uncertainties

A. Optimal Joint Operation Model of IEHSs

1) Multiple Objectives for Multiple Operators

In the optimal joint operation model of IEHSs, the hydraulic thermal flow in each isolated DHS is coupled with the power flow in the adjacent distribution sub-network via a CHP plant. The joint operation aims to achieve social benefits by minimizing multiple objectives for PDS and DHS operators as follows.

$$\min \left\{ \sum_{k \in K} f_k^H, \sum_{m \in M} f_m^P \right\} \tag{9}$$

where K is the set of DHSs; m and M are the set and index of distribution sub-networks, respectively; f_k^H denotes the optimization objective for each DHS; and f_m^P denotes the optimization for each distribution sub-network.

The optimal joint operation model of IEHSs can be decoupled into two optimization models for DHS and PDS, respectively, which are introduced as follows.

2) DHS Optimization Model

The objective of DHS optimization model is to minimize the total thermal network loss and the penalty of temperature constraint violation. This is equivalent to the minimization of the total operation cost while fulfilling customer requirements. The DHS optimization model can be formulated as:

$$\begin{cases} \min\limits_{\phi_{k,t}^{CHP}} f_k^H = \sum\limits_{t \in OT} \left(\phi_{k,t}^{loss} + \lambda^T \varpi_{k,t}^{T_{viol}} \right) \\ \text{s.t. (1)-(8)} \end{cases} \tag{10}$$

$$\varpi_{k,t}^{T_{viol}} = \sum_{n \in N_k} \max\left\{ \left(T_{n,t}^{sup} - T_n^{sup,\max} \right), 0, \left(T_n^{sup,\min} - T_{n,t}^{sup} \right) \right\} + \max\left\{ \left(T_{k,t}^{src} - T_k^{src,\max} \right), 0, \left(T_k^{src,\min} - T_{k,t}^{src} \right) \right\} \tag{11}$$

where $\phi_{k,t}^{loss}$ is the thermal power loss of DHS k; λ^T is the penalty factor for temperature; $\varpi_{k,t}^{T_{viol}}$ is the temperature violation of DHS k; OT is the set of timesteps; and N_k is the set of nodes of DHS k. The penalty of temperature constraint violation in (10) is calculated as the summation of all water temperature deviations out of the allowed ranges in (11).

3) PDS Optimization Model

The objective of PDS optimization model is to minimize the total network power loss and the penalty of voltage constraint violation. A PDS optimization model can be formulated as:

$$\min_{p_{i,t}^{RDG}, q_{i,t}^{RDG}} f_m^P = \sum_{t \in OT} \left(p_{m,t}^{loss} + \lambda^V \varpi_{m,t}^{V_{viol}} \right) \tag{12}$$

s.t.

$$p_{i,t}^{CHP}=\frac{1}{\eta}\phi_{i,t}^{CHP}\quad\forall i \tag{13}$$

$$\left(p_{i,t}^{RDG}\right)^2+\left(q_{i,t}^{RDG}\right)^2\le\left(S_i^{RDG}\right)^2\quad\forall i \tag{14}$$

$$0\le p_{i,t}^{RDG}\le p_{i,t}^{RDG,\max}\quad\forall i \tag{15}$$

$$-\delta S_i^{RDG}\le q_{i,t}^{RDG}\le\delta S_i^{RDG}\quad\forall i \tag{16}$$

$$\varpi_{m,t}^{V_{viol}}=\sum_{i\in N_m}\max\left\{\left(v_{i,t}-v_i^{\max}\right),0,\left(v_i^{\min}-v_{i,t}\right)\right\}\quad\forall i \tag{17}$$

$$p_{i,t}^{RDG}+p_{k,t}^{CHP}-p_{i,t}^{d}=\sum_{j\in N_m}\left|v_{i,t}\right|\left|v_{j,t}\right|\left(G_{ij}\cos\left(\vartheta_{i,t}-\vartheta_{j,t}\right)+B_{ij}\sin\left(\vartheta_{i,t}-\vartheta_{j,t}\right)\right)\quad\forall i \tag{18}$$

$$q_{i,t}^{RDG}-q_{i,t}^{d}=\sum_{j\in N_m}\left|v_{i,t}\right|\left|v_{j,t}\right|\left(G_{ij}\sin\left(\vartheta_{i,t}-\vartheta_{j,t}\right)-B_{ij}\cos\left(\vartheta_{i,t}-\vartheta_{j,t}\right)\right)\quad\forall i \tag{19}$$

where ij is the index of branches in distribution sub-network m; i and j are the indexes of buses in distribution sub-network m; $p_{m,t}^{loss}$ is the active power loss of distribution sub-network m; G_{ij} and B_{ij} are the branch conductance and susceptance, respectively; λ^V is the penalty factor for voltage violations; N_m is the set of buses in distribution sub-network m; $p_{i,t}^{CHP}$ is the electrical power generation of CHP plant; $p_{i,t}^{d}$ and $q_{i,t}^{d}$ are the active and reactive power of electrical load, respectively; $p_{i,t}^{RDG}$ and $q_{i,t}^{RDG}$ are the RDG active and reactive power, respectively; $p_{i,t}^{RDG,\max}$ is the maximum available RDG active power; $\varpi_{m,t}^{V_{viol}}$ is the voltage violation in distribution sub-network m; δ is the reactive power capacity factor of the inverter; η is the power transfer efficiency factor; S_i^{RDG} is the apparent power capacity of the installed RDG; $v_{i,t}$ and $\vartheta_{i,t}$ are the bus voltage magnitude and angle, respectively; and $v_i^{\max}$ and $v_i^{\min}$ are the maximum and minimum voltage limits, respectively.

With the thermal power outputs of CHP plant given by DHSs, the electrical power outputs of CHP plant can be calculated by (13). Constraints (14)-(16) represent a practical operation limits on inverter-based RDGs with the var-priority mode, where the active power could be curtailed only if the reactive power is insufficient. The penalty of voltage constraint violation is calculated as the summation of all bus voltage deviations out of the allowed ranges by (17). The AC power flow model is presented by (18) and (19), and it can be solved by the Newton-Raphson iterative algorithm. The decision variables of the optimization model are $\left\{T^{sup},T^{src},P^{CHP},\phi^{CHP},P^{RDG}q^{RDG},|v|,\theta\right\}$.

The above optimization models of DHS and PDS are computationally expensive. Due to the coupled model of CHP plant in (13), one of the separate models of DHS and PDS with corresponding objectives could be infeasible when the coupled power outputs are determined and fixed by the other one. This is one of the motivations that the application of DRL method is highly expected. The hydraulic-thermal flow and power flow algorithms can be conducted to provide simulation environment for DRL training process.

B. Stochastic Programming

It is essential to consider uncertainties of RDG active power generation, active and reactive power consumptions, and thermal demands in the optimal joint operation model of IEHSs. Locational variations of RDG generation and electrical and thermal load predictions are regarded as spatial uncertainties, while short-term data errors due to power intermittency and communication/response delays are considered as temporal uncertainties. We apply a stochastic programming method to consider the spatial and temporal uncertainties via random scenarios for robust operation against uncertainty realizations.

In both PDS and DHS, system operation conditions may vary geographically across the entire systems. These variations depend on uncertain user patterns, RDG operation status, etc. The optimal joint operation model of IEHSs is based on predicted intervals of RDG active power generation, active and reactive power consumptions, and thermal demands. In particular, $p_{i,t_0}^{RDG,\max}$, p_{i,t_0}^{d}, q_{i,t_0}^{d}, and ϕ_{n,t_0}^{d} indicate spatial and temporal uncertainties at t_0, constrained by the intervals as below.

$$\underline{P}_{i,\zeta}^{RDG}\le p_{i,t_0}^{RDG,\max}\le\overline{P}_{i,\zeta}^{RDG}\quad\forall i \tag{20}$$

$$\underline{P}_{i,\zeta}^{d}\le p_{i,t_0}^{d}\le\overline{P}_{i,\zeta}^{d}\quad\forall i \tag{21}$$

$$\underline{Q}_{i,\zeta}^{d}\le q_{i,t_0}^{d}\le\overline{Q}_{i,\zeta}^{d}\quad\forall i \tag{22}$$

$$\underline{\Phi}_{n,\zeta}^{d}\le\phi_{n,t_0}^{d}\le\overline{\Phi}_{n,\zeta}^{d}\quad\forall n \tag{23}$$

where $\underline{P}_{i,\zeta}^{RDG}$ and $\overline{P}_{i,\zeta}^{RDG}$ are the spatial uncertainty and temporal uncertainty of the maximum power points of RDG, respectively; $\underline{P}_{i,\zeta}^{d}$ and $\overline{P}_{i,\zeta}^{d}$ are the spatial uncertainty and temporal uncertainty of active power loads, respectively; $\underline{Q}_{i,\zeta}^{d}$ and $\overline{Q}_{i,\zeta}^{d}$ are the spatial uncertainty and temporal uncertainty of reactive power loads, respectively; and $\underline{\Phi}_{n,\zeta}^{d}$ and $\overline{\Phi}_{n,\zeta}^{d}$ are the upper and lower bounds of spatial uncertainty of thermal power consumption, respectively. To further eliminate the impacts of real-time power intermittency and communication/response delays, temporal uncertainties during each timestep are formulated as below. The intervals in this temporal uncertainty model vary temporally around the real-time measurements at t.

$$\underline{P}_{i,\tau}^{RDG}\le p_{i,t}^{RDG,\max}\le\overline{P}_{i,\tau}^{RDG}\quad\forall i,t\ne t_0 \tag{24}$$

$$\underline{P}_{i,\tau}^{d}\le p_{i,t}^{d}\le\overline{P}_{i,\tau}^{d}\quad\forall i,t\ne t_0 \tag{25}$$

$$\underline{Q}_{i,\tau}^{d}\le q_{i,t}^{d}\le\overline{Q}_{i,\tau}^{d}\quad\forall i,t\ne t_0 \tag{26}$$

Random scenarios are generated for spatial and temporal uncertainties. With sufficient scenarios, a scenario-based stochastic optimization model is formulated to obtain robust solutions. Note that the scenario-based stochastic optimization model is highly time consuming, and it is impossible for them to handle temporal uncertainties within a required short computation time. This is another motivation for the application of DRL method.

V. Application of MADDPG Method

The optimal joint operation model of IEHSs contains the thermal and electrical power of CHP plants. The hydraulic-thermal and power flows could be solved alternatively for a joint operation decision. However, the non-linear and non-convex models lead to extensive computation time, and temporal uncertainties could not be addressed by conventional programming methods. To solve these issues, we develop a multi-agent DRL method with MADDPG method.

A. DRL-based Optimization Model

The proposed optimal joint operation model of IEHSs is reformulated into a DRL-based optimization model. In this formulation, each control center of distribution sub-network and DHS is modeled as an agent that suits the DRL training and execution requirements.

For the PDS agent m, an observation space $\mathcal{O}_m^P$ covers the dynamic power system information including $v_{i,t}$, $p_{i,t}^d$, $q_{i,t}^d$, and $p_{i,t}^{RDG,\max}$ within the corresponding distribution sub-network. An action space $\mathcal{A}_m^P$ includes the continuous active and reactive power dispatch signals for each RDG inside the distribution sub-network, i.e., $p_{i,t}^{RDG}$ and $q_{i,t}^{RDG}$. A reward function is modeled based on (12). In the DRL method, the reward to be maximized, i.e., $r_{m,t}^P$, is formulated as:

$$r_{m,t}^P = -\left(p_{m,t}^{loss} + \lambda^V \varpi_{m,t}^{V_{viol}}\right) \tag{27}$$

An observation space of the DHS agent k, $\mathcal{O}_k^H$, covers the dynamic DHS information including $\phi_{n,t}^d$, $T_{k,t}^{src}$, and $T_{n,t}^{sup}$. An action space $\mathcal{A}_k^H$ includes $\phi_{k,t}^{CHP}$. These actions are generated while ensuring that the thermal power outputs are larger than the loads. Similarly, a reward function $r_{k,t}^H$ is modeled based on (10) as:

$$r_{k,t}^H = -\left(\phi_{k,t}^{loss} + \lambda^T \varpi_{k,t}^{T_{viol}}\right) \tag{28}$$

A transition function $\mathbb{T}$ represents the system dynamics. It takes the environment to the next state after all agents adopt actions in the current state. In this paper, it directly corresponds to the optimal joint operation model of IEHSs. The observation spaces form an environment state space, i.e., $\mathbb{S} = \mathcal{O}_1^P \cup \ldots \cup \mathcal{O}_M^P \cup \mathcal{O}_1^H \cup \ldots \cup \mathcal{O}_k^H$. Together with the action space, reward function, and transition function, the partially observable Markov game model presented as $\{\mathbb{S}, \mathbb{A}, \mathbb{R}, \mathbb{T}\}$ is then formulated, where each agent takes actions only based on its own observation and a respective actor network, $\boldsymbol{\pi}_m^P$ or $\boldsymbol{\pi}_k^D$. With the actions of all agents taken in the current environment state, the system is taken into the next state through the transition function $\mathbb{T}: \mathcal{S} \times a_1^P \times \ldots \times a_M^P \times a_1^H \times \ldots \times a_K^H \to \mathcal{S}'$. The target of each agent is to maximize its overall expected return R defined as:

$$R = \sum_{t \in OT} \gamma^t r_t \tag{29}$$

where γ^t is the reward discount factor.

Given a state s and a state-action pair (s, a), the performance evaluation is derived as a state value function $V^\pi(s)$ and a state-action value function $Q^\pi(s, a)$, respectively. Generally, in DRL problems, these functions are established on the expected, discounted, and accumulated return on the state s and state-action pair (s, a) [29]. The state value function and the state-action value function are formulated as:

$$\begin{cases} V^\pi(s) = \mathbb{E}_\pi\left\{R_t \middle| s_t = s\right\} = \mathbb{E}_\pi\left\{\sum_{k=0}^{OT-t} \gamma^k r_{t+k+1} \middle| s_t = s\right\} \\ Q^\pi(s, a) = \mathbb{E}_\pi\left\{R_t \middle| s_t = s, a_t = a\right\} = \mathbb{E}_\pi\left\{\sum_{k=0}^{OT-t} \gamma^k r_{t+k} \middle| s_t = s, a_t = a\right\} \end{cases} \tag{30}$$

where $\mathbb{E}$ represents the expectation function.

Both functions map each agent from the current state to its potential final performance. The state-action value function is mainly discussed and applied in the DRL model. When deterministic policies are given, the state-action value function (30) can be reformulated with the Bellman equation [29] as:

$$Q^{\pi^*}(s, a) = r(s, a, s') + \gamma \max_{a'} Q^{\pi^*}(s', a') \tag{31}$$

where π^* represents the optimal policy, mapping from the current state $\mathcal{S}$ to the action $\mathcal{A}$ that the agent should take.

B. MADDPG Method

In an MADDPG method, agents are deployed to learn the optimal policies via a centralized training manner. Then in the decentralized execution, the optimal actions of each agent can be provided according to local observations. There are no specific communication links required among agents, and this method can be applied for a cooperative environment [30]. Thus, an MADDPG method is developed to solve the optimal joint operation problem of IEHSs. Note that the central control center takes the responsibility of the centralized training, while the local control centers work as the agents to make online decisions.

Each agent is equipped with a set of neural networks including actor networks, critic networks, target actor networks, and target critic networks. The parameters of these networks are denoted as θ^π, θ^Q, $\theta^{\pi'}$, and $\theta^{Q'}$, respectively. The actor networks are mathematical approximations of agent policies to take actions based on given observations, i.e., $\boldsymbol{\pi}(o_t|\theta^\pi)$. On the other hand, the critic networks are mathematical approximations for the centralized state-action value function given in (31), i. e., $\boldsymbol{Q}(s, a|\theta^Q)$, with observations and actions from all agents, $s = \{o_1^P, o_2^P, \ldots, o_M^P, o_1^H, o_2^H, \ldots, o_K^H\}$ and $a = \{a_1^P, a_2^P, \ldots, a_M^P, a_1^H, a_2^H, \ldots, a_K^H\}$. The target actor networks and target critic networks are deployed in the MADDPG method to train the agents towards maximizing its overall expected return formulated as (29).

The purpose of the training stage is for each agent to learn its optimal policy $\boldsymbol{\pi}^*$. To achieve this, θ^π is updated towards maximizing the performance objective $J(\theta^\pi)$, which means in the direction of $\nabla_{\theta^\pi} J(\theta^\pi)$. $\nabla_{\theta^\pi} J(\theta^\pi)$ is often denoted as the policy gradient, and it is approximated as:

$$\nabla_{\theta^\pi} J(\theta^\pi) \approx \mathbb{E}_{\boldsymbol{s} \sim \rho^{\boldsymbol{\Pi}}, \boldsymbol{a} \sim \theta^\pi}\left[\nabla_{\theta^\pi} \ln \boldsymbol{\pi}(o|\theta^\pi) \boldsymbol{Q}(s, a|\theta^Q)\right] \tag{32}$$

where $\rho^{\boldsymbol{\Pi}}$ denotes the state distribution; and $\boldsymbol{s} \sim \rho^{\boldsymbol{\Pi}}$ and $\boldsymbol{a} \sim \theta^\pi$ denote the states sampled from $\rho^{\boldsymbol{\Pi}}$ and the actions sampled from $\boldsymbol{\pi}$ in the state s, respectively. Details can be found in [31].

Once the neural networks are initialized, all agents start interaction and exploration with the environment. For the spatial uncertainties, a stochastic state has been applied for each episode during the exploration. These states are generated based on (20)-(23). Within each timestep, the temporal uncertainties are sampled according to (24)-(26). In specific, the actions are taken into these parallel stochastic scenarios, and the rewards are derived according to the optimal joint operation information of IEHSs in all scenarios. The transitions are then stored into a prioritized experience replay with flash replay [17], which improves the sampling efficiency.

The detailed procedure of the MADDPG method is explained in Supplementary Material A. With enough accumulation of experience replay, the networks of each agent are updated. Moreover, the Q-loss functions are derived with temporal difference errors, and they are utilized for the critic network update, where the policy gradients are further derived and utilized for the actor network update. Then, the corresponding target networks are updated via a soft update velocity factor $\acute{o}$.

C. Overall Implementation Process

The overall implementation process of optimal joint operation method of IEHSs based on multi-agent DRL method is demonstrated in Fig. 3. In this process, the PDS agents and DHS agents accumulate experience through interactions with the environment of combined power flow and hydraulic-thermal flow. Then, they learn to improve all rewards via the neural network training process given in Supplementary Material A Algorithm SA1. Note that due to IEHS coupling, the thermal power generation from the action space is also used in the state space for PDS agents to train their neural networks.

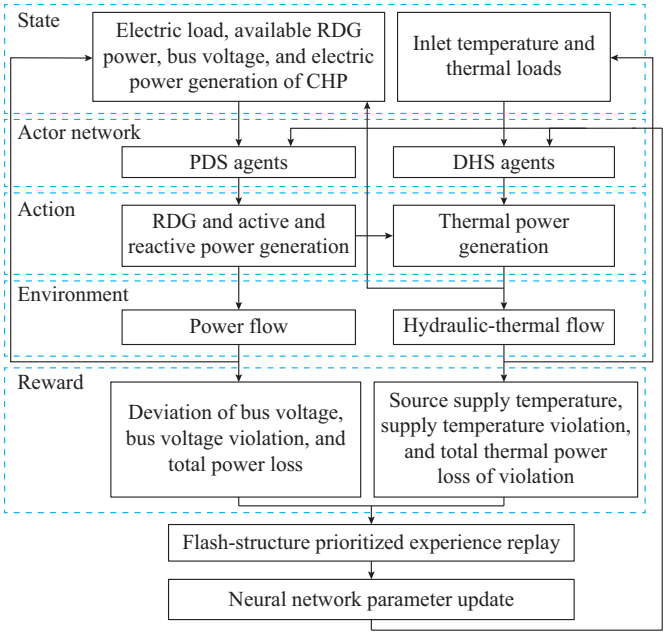

Fig. 3. Overall implementation process of optimal joint operation method of IEHSs based on multi-agent DRL method.

The spatial uncertainties are fully modeled as scenarios in the system state initialization, while the temporal uncertainties are fully modeled during each interaction with the environment. The MADDPG method is specifically developed to include the developed hydraulic-thermal flow model in the training process, thus solving the optimal joint operation problem of IEHSs.

VI. Case Study

A. Test System Description

Numerical simulation tests are conducted on an IEHS with the standard 33-bus PDS [32], three practical DHSs [3], and assuming 120% scale thermal loads. The PDS consists of 3 interconnected distribution sub-networks, each coupled with a DHS via a CHP plant unit. The topology of test IEHS is displayed in Fig. 4, where T is short for transformer. RDGs are installed at 12 buses as shown with circles, where the capacities are same with the electric loads of the local bus, with $\delta = 0.6$. The base power is 10 MVA. The allowed bus voltage range is set to be [0.95 p.u., 1.05 p.u.] for all buses. λ^V is set to be 10. In the DHSs, the supply temperature range is set to be $[70, 90]$℃. The return temperature is initialized as 30 ℃. Without the loss of generality, the electric power capacity of CHP plants is 1 MW and $\eta = 78.99\%$. The above system settings are designed according to industrial projects of active distribution networks [33] and DHSs [34]. The parameters for the test IEHS and loads are recorded and uploaded online, which can be found in [35]. The household power and thermal loads are aggregated to the bus and node level, which is a main assumption. Note that, when necessary, other test systems can be applied without affecting the effectiveness of the proposed method which is generic to any IEHS settings.

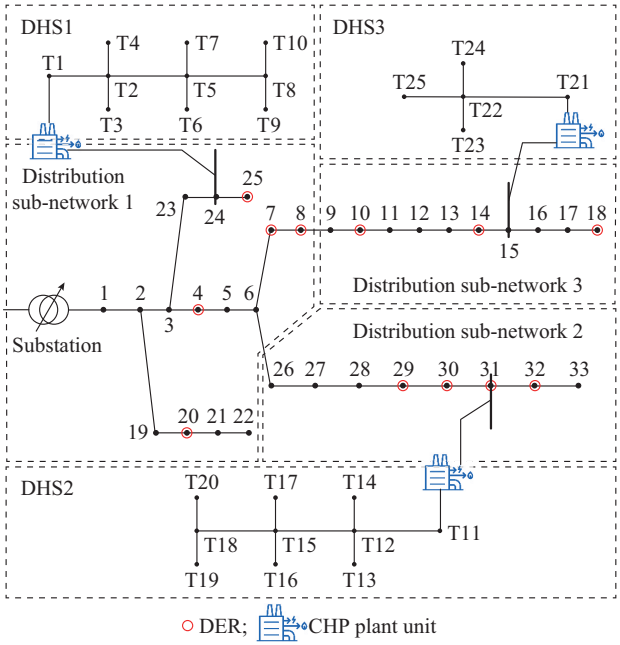

Fig. 4. Topology of test IEHS.

The Adam optimizer is applied [36], with learning rates of 10^{-5}, $\gamma = 0.9$, and $\acute{o} = 0.01$. All rewards in the algorithm are scaled up by 10^3. The target Q-value estimation factor $\mu =$

0.95. The capacities of the main experience replay and the flash experience replay are 2×10^4 and 10^3, respectively. The minibatch size is 128. $\varepsilon=6$ and it starts to decay by 0.1% towards 0 at every timestep during the training. For comparison, a widely used on-policy PPO method [37] is applied to the same test system. For the clipped surrogate objective, we define ε to be 0.1, and γ to be 0.8. The learning rates of actor and critic networks are both 10^{-5}.

The simulations are conducted on a 64-bit PC with an Intel 8-core 3.60 GHz i7-9700K CPU and one NVIDIA GeForce RTX 2080 Ti GPU using Python platform, with the AC power flow solved by the PYPOWER solver [38].

Two typical cases are designed for numerical simulations.

1) Case 1: high-RDG low-load case with 100% RDG power outputs, 50% electric loads, and 80% thermal loads.

2) Case 2: low-RDG high-load case with 0% RDG power outputs, 100% electric loads, and 100% thermal loads.

The uncertainty setup of the two cases is given in Table I.

TABLE I
UNCERTAINTY SETUP OF TWO CASES

Case	$\left[\underline{P}_{i,\zeta}^{RDG},\overline{P}_{i,\zeta}^{RDG}\right]$	$\left[\underline{P}_{i,\zeta}^{d},\overline{P}_{i,\zeta}^{d}\right]$	$\left[\underline{Q}_{i,\zeta}^{d},\overline{Q}_{i,\zeta}^{d}\right]$	$\left[\underline{\Phi}_{n,\zeta}^{d},\overline{\Phi}_{n,\zeta}^{d}\right]$	$\left[\underline{P}_{i,\tau}^{RDG},\overline{P}_{i,\tau}^{RDG}\right]$	$\left[\underline{P}_{i,\tau}^{d},\overline{P}_{i,\tau}^{d}\right]$	$\left[\underline{Q}_{i,\tau}^{d},\overline{Q}_{i,\tau}^{d}\right]$
1	$[0.7,1.0]$	$[0.4,0.6]$	$[0.4,0.6]$	$[0.7,0.9]$	$[0.75,1.25]\,p_{i,t}^{RDG,\max}$	$[0.95,1.05]\,p_{i,t}^{d}$	$[0.75,1.25]q_{i,t}^{d}$
2	$[0,0]$	$[0.8,1.0]$	$[0.8,1.0]$	$[0.9,1.1]$			

B. Convergence of Data-driven Optimization

The convergences of the proposed method in cases 1 and 2 are demonstrated in Figs. 5 and 6, respectively. The results demonstrate the overall good performance and coordination between PDS and DHS agents. The training results of PDS agent show good convergence avoiding voltage violations and reducing power losses. Meanwhile, the DHS agents have avoided temperature violations and reduced thermal power losses after training.

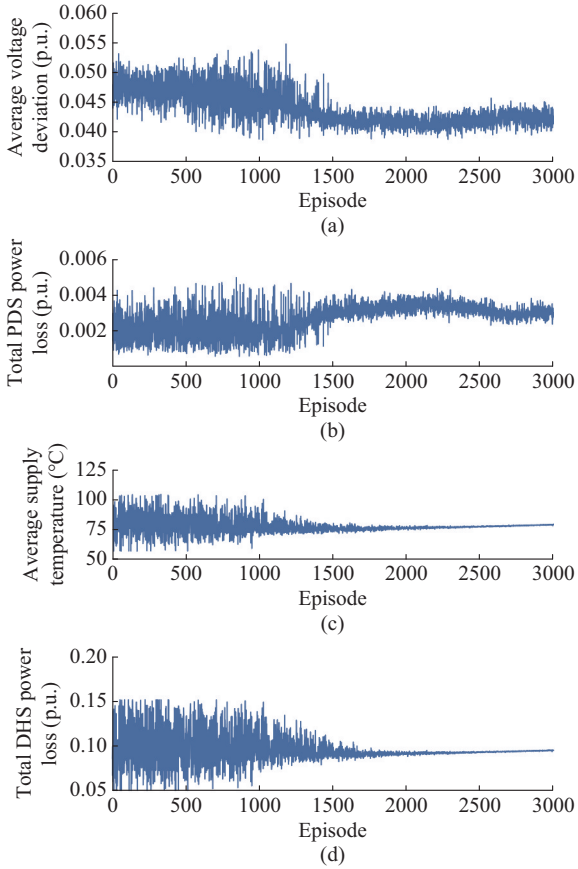

Fig. 5. Convergence of proposed method in case 1. (a) Average voltage deviation. (b) Total PDS power loss. (c) Average supply temperature. (d) Total DHS power loss.

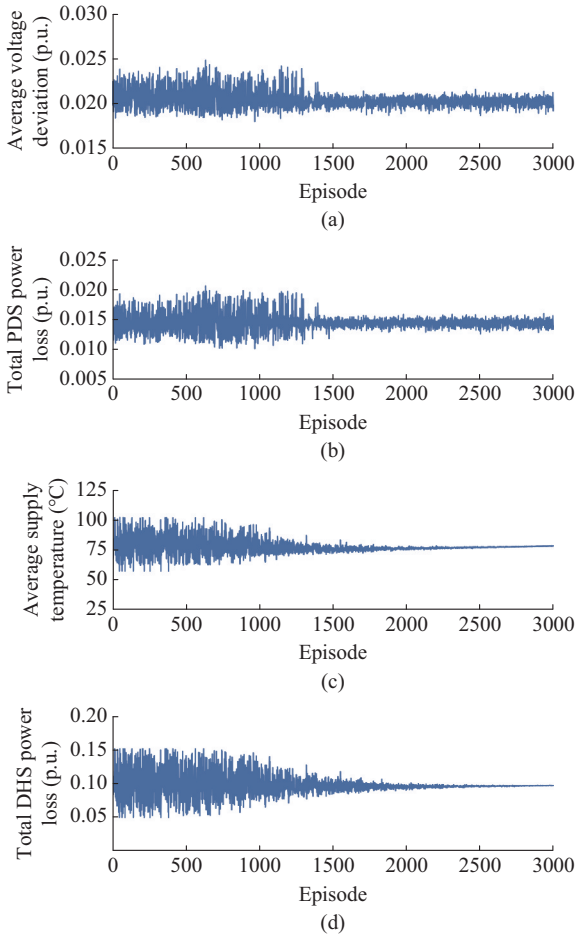

Fig. 6. Convergence of proposed method in case 2. (a) Average voltage deviation. (b) Total PDS power loss. (c) Average supply temperature. (d) Total DHS power loss.

The convergences of the PPO method during the training process in cases 1 and 2 are illustrated in Figs. 7 and 8, respectively. Compared with PPO method, the proposed method outperforms on the convergence of supply temperature and DHS power loss in cases 1 and 2, with the smaller training fluctuation in episode 3000.

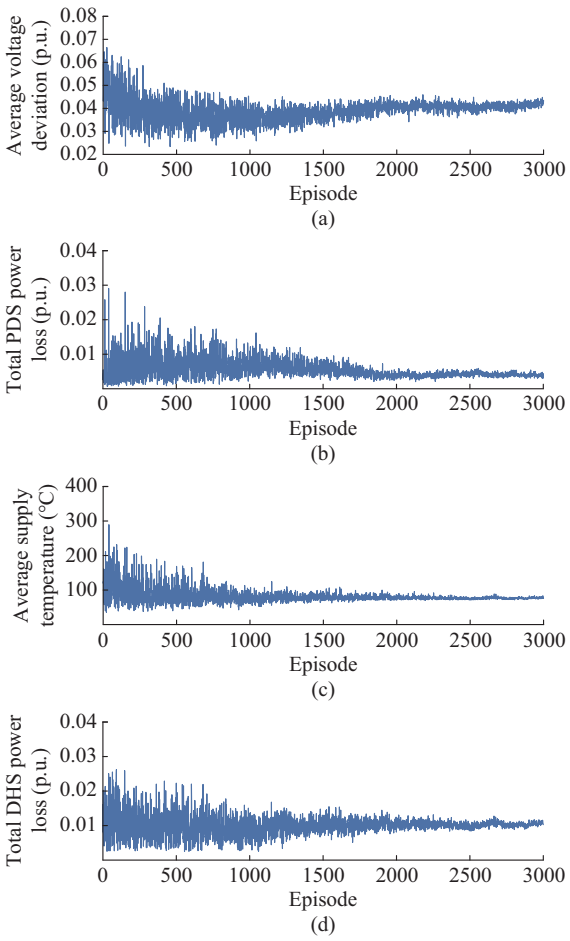

Fig. 7. Convergence of PPO method during training process in case 1. (a) Average voltage deviation. (b) Total PDS power loss. (c) Average supply temperature. (d) Total DHS power loss.

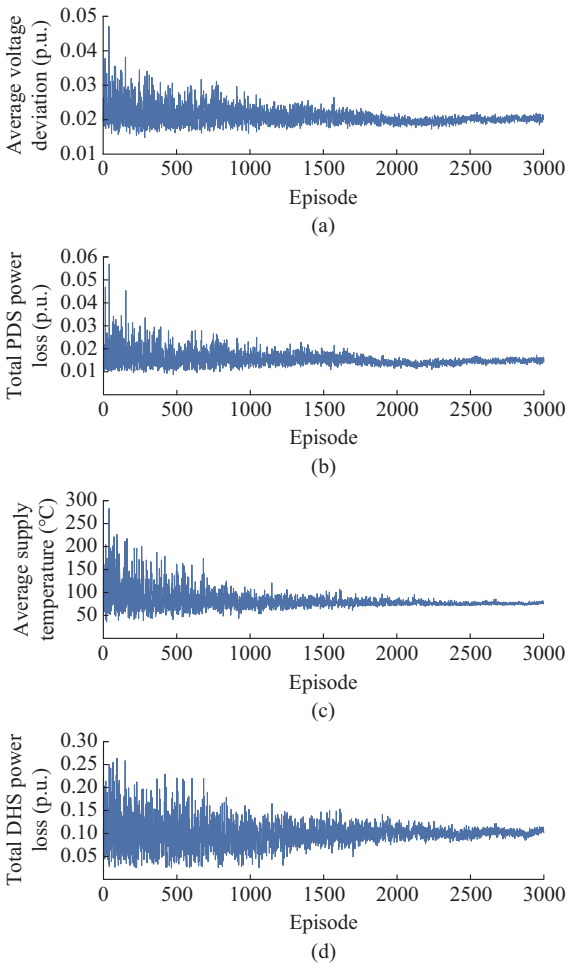

Fig. 8. Convergence of PPO method during training process in case 2. (a) Average voltage deviation. (b) Total PDS power loss. (c) Average supply temperature. (d) Total DHS power loss.

C. Optimization Results

To validate the effectiveness of the proposed method in real time, 100 Monte-Carlo scenarios in each case are sampled and used as testing data, based on the setup in Table I. The results of the proposed method in two cases are shown in Table II. It shows that the proposed method performs well after training, without voltage violations, significantly reducing voltage deviations and network power losses.

The temperature violation is eliminated with thermal power loss effectively reduced in both cases. The results of PPO method under two cases are shown in Table III. In comparison with the PPO method, the proposed method outperforms with the lower maximum voltage deviation, average total PDS power loss, and average total DHS power loss in both cases.

TABLE II
RESULTS OF PROPOSED METHOD IN TWO CASES

Case	The maximum voltage deviation (p.u.)	Average voltage deviation (p.u.)	Number of voltage violations	Average total PDS power loss (p.u.)	The maximum temperature (°C)	Average temperature (°C)	Number of supply temperature violations	Average total DHS power loss (p.u.)
Case 1	0.0498	0.0422	0	0.0029	88.15	77.32	0	0.0149
Case 2	0.0475	0.0201	0	0.0143	82.70	76.53	0	0.0153

TABLE III
RESULTS OF PPO METHOD IN TWO CASES

Case	The maximum voltage deviation (p.u.)	Average voltage deviation (p.u.)	Number of voltage violations	Average total PDS power loss (p.u.)	The maximum temperature (°C)	Average temperature (°C)	Number of supply temperature violations	Average total DHS power loss (p.u.)
Case 1	0.0499	0.0420	0	0.0038	80.97	75.97	0	0.017
Case 2	0.0476	0.0204	0	0.0150	81.15	76.68	0	0.016

D. Validation of Solution Robustness

To validate the solution robustness of the proposed method, additional 5000 random scenarios have been generated as spatial and temporal uncertainties. For a comparison purpose, a deterministic optimization (DO) method and PPO method are applied with the same settings. Both methods are comparable to the proposed method for optimization and computation efficiency. The results of different methods in cases 1 and 2 are shown in Tables IV and V, respectively. In addition to the voltage violations and average values, the absolute voltage deviation and total PDS power loss (with 95th percentiles) are analyzed and compared. The 95th percentiles indicate the value at risk at the confidential level of 95%.

TABLE IV
RESULTS OF DIFFERENT METHODS IN CASE 1

Method	Number of voltage violations	Average absolute voltage deviation (p.u.)	Absolute voltage deviation with 95th percentiles (p.u.)	Average total PDS power loss (kW)	Total PDS power loss with 95th percentiles (kW)
DO	14	0.0382	0.0497	26.12	35.49
PPO	1	0.0421	0.0498	37.75	43.8
Proposed	0	0.0423	0.0498	29.14	33.45

TABLE V
RESULTS OF DIFFERENT METHODS IN CASE 2

Method	Number of voltage violations	Average absolute voltage deviation (p.u.)	Absolute voltage deviation with 95th percentiles (p.u.)	Average total PDS power loss (kW)	Total PDS power loss with 95th percentiles (kW)
DO	0	0.0259	0.0483	140.3	158.09
PPO	0	0.0204	0.0476	149.5	159.70
Proposed	0	0.0201	0.0475	142.8	152.17

The proposed method demonstrates full robustness against temporal and spatial uncertainties, where no voltage magnitudes are found that violate the voltage limits among 5000 random scenarios in both cases. However, in case 1, DO and PPO methods lead to 14 and 1 out of 5000 random scenarios with voltage violations during the operation, respectively. In case 2, although DO and PPO methods make no voltage violation, the proposed method significantly outperforms in minimizing the voltage deviations with the similar or reduced power losses, showing better optimization efficiency. Since the proposed method considers stochastic samples in training neural networks, it can better control the risk on the objective of power loss reduction. It is evidenced that the proposed method can achieve the less total PDS power loss with 95th percentiles than the DO method, although the values of average total PDS power loss are a bit higher. The execution time of the proposed method and PPO method is 0.55 s and 0.54 s on average in cases 1 and 2, respectively.

VII. CONCLUSION

In summary, an optimal joint operation method of IEHSs based on multi-agent DRL method is proposed in this paper, minimizing multiple objectives while satisfying operation constraints. An MADDPG method is developed with a new compatible hydraulic-thermal flow algorithm. The temporal and spatial uncertainties have been fully considered and addressed by the proposed method. Through the case studies, the proposed method has shown great performance after the efficient training procedure. Besides, it also outperforms a conventional DO method in two cases, with robust minimization results of voltage deviations and power losses.

In future works, advanced DRL training techniques including imitation learning and prioritized experience replay, as well as advanced DRL methods such as twin-delayed DDPG, can be utilized to further improve the training efficiency and apply to a larger-scale IEHS. To guarantee no violations of operation constraints, safe action space techniques can be further introduced. Furthermore, the optimization of larger energy systems or energy clusters can be studied with an application of the proposed method. Advanced learning structures such as federated learning and curriculum learning can be applied to further enhance the efficiency of the proposed method. Moreover, the potential of utilizing electric vehicle overnight charging strategies and vehicle-to-grid functions to improve the economic and technical benefits for IEHSs will be investigated.

REFERENCES

[1] W. Zheng, Y. Hou, and Z. Li, "A dynamic equivalent model for district heating networks: formulation, existence and application in distributed electricity-heat operation," *IEEE Transactions on Smart Grid*, vol. 12, no. 3, pp. 2685-2695, May 2021.
[2] S. Lu, W. Gu, K. Meng *et al*., "Thermal inertial aggregation model for integrated energy systems," *IEEE Transactions on Power Systems*, vol. 35, no. 3, pp. 2374-2387, May 2020.
[3] X. Liu, J. Wu, N. Jenkins *et al*., "Combined analysis of electricity and heat networks," *Applied Energy*, vol. 162, pp. 1238-1250, Jan. 2016.
[4] X. Qin, H. Sun, X. Shen *et al*., "A generalized quasi-dynamic model for electric-heat coupling integrated energy system with distributed energy resources," *Applied Energy*, vol. 251, p. 113270, Oct. 2019.
[5] L. Mitridati and P. Pinson, "Optimal coupling of heat and electricity systems: a stochastic hierarchical approach," in *Proceedings of 2016 International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*, Beijing, China, Jun. 2016, pp. 1-6.
[6] A. Hellmers, M. Zugno, A. Skajaa *et al*., "Operational strategies for a portfolio of wind farms and CHP plants in a two-price balancing market," *IEEE Transactions on Power Systems*, vol. 31, no. 3, pp. 2182-2191, May 2016.
[7] H. Cai, S. You, J. Wang *et al*., "Technical assessment of electric heat boosters in low-temperature district heating based on combined heat and power analysis," *Energy*, vol. 150, pp. 938-949, May 2018.
[8] Z. Li, W. Wu, M. Shahidehpour *et al*., "Combined heat and power dis-

patch considering pipeline energy storage of district heating network," in *Proceedings of 2017 IEEE PES General Meeting*, Chicago, USA, Jul. 2017, pp. 1-8.
[9] C. Shao, Y. Ding, J. Wang *et al*., "Modeling and integration of flexible demand in heat and electricity integrated energy system," *IEEE Transactions on Sustainable Energy*, vol. 9, no. 1, pp. 361-370, Jan. 2018.
[10] S. Lu, W. Gu, C. Zhang *et al*., "Hydraulic-thermal cooperative optimization of integrated energy systems: a convex optimization approach," *IEEE Transactions on Smart Grid*, vol. 11, no. 6, pp. 4818-4832, Nov. 2020.
[11] H. Liu. (2023, Aug.). Deep reinforcement learning for distribution network operation and electricity market. [Online]. Available: https://unsworks. unsw. edu. au/entities/publication/ab444d83-f392-4bda-9209-6706ddd16
[12] Y. Zhang, Y. Xu, Y. Mishra *et al*., "A master-slave deep learning framework for real-time transient stability-constrained optimal power flow," *IEEE Transactions on Power Systems*, vol. 40, no. 3, pp. 2674-2687, May 2025.
[13] Q. Ma and C. Deng, "Deterministic and robust volt-var control methods of power system based on convex deep learning," *Journal of Modern Power Systems and Clean Energy*, vol. 12, no. 3, pp. 719-729, May 2024.
[14] Y. Wang, D. Qiu, and G. Strbac, "Multi-agent deep reinforcement learning for resilience-driven routing and scheduling of mobile energy storage systems," *Applied Energy*, vol. 310, p. 118575, Mar. 2022.
[15] M. Shin, D. H. Choi, and J. Kim, "Cooperative management for PV/ESS-enabled electric vehicle charging stations: a multiagent deep reinforcement learning approach," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 5, pp. 3493-3503, May 2020.
[16] W. Wang, N. Yu, Y. Gao *et al*., "Safe off-policy deep reinforcement learning algorithm for volt-var control in power distribution systems," *IEEE Transactions on Smart Grid*, vol. 11, no. 4, pp. 3008-3018, Jul. 2020.
[17] H. Liu, C. Zhang, Q. Chai *et al*., "Robust regional coordination of inverter-based volt/var control via multi-agent deep reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 12, no. 6, pp. 5420-5433, Nov. 2021.
[18] X. Chen, G. Qu, Y. Tang *et al*., "Reinforcement learning for selective key applications in power systems: recent advances and future challenges," *IEEE Transactions on Smart Grid*, vol. 13, no. 4, pp. 2935-2958, Jul. 2022.
[19] C. Huang, H. Zhang, L. Wang *et al*., "Mixed deep reinforcement learning considering discrete-continuous hybrid action space for smart home energy management," *Journal of Modern Power Systems and Clean Energy*, vol. 10, no. 3, pp. 743-754, May 2022.
[20] J. Wang, Y. Wang, D. Qiu *et al*., "Resilient energy management of a multi-energy building under low-temperature district heating: a deep reinforcement learning approach," *Applied Energy*, vol. 378, p. 124780, Jan. 2025.
[21] B. Wang, C. Zhang, X. Chen *et al*., "Price-based demand response supported three-stage hierarchically coordinated voltage control for microgrids," *Journal of Modern Power Systems and Clean Energy*, vol. 13, no. 1, pp. 338-350, Jan. 2024.
[22] R. Lu, Z. Jiang, H. Wu *et al*., "Reward shaping-based actor-critic deep reinforcement learning for residential energy management," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 3, pp. 2662-2673, Mar. 2023.
[23] 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.
[24] H. Shengren, E. M. Salazar, P. P. Vergara *et al*. (2022, Aug.). Performance comparison of deep RL algorithms for energy systems optimal scheduling. [Online]. Available: https://arxiv.org/abs/2208.00728
[25] 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.
[26] S. Li, D. Cao, W. Hu *et al*., "Multi-energy management of interconnected multi-microgrid system using multi-agent deep reinforcement learning," *Journal of Modern Power Systems and Clean Energy*, vol. 11, no. 4, pp. 1606-1617, Jul. 2023.
[27] D. Qiu, T. Chen, G. Strbac *et al*., "Coordination for multienergy microgrids using multiagent reinforcement learning," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 4, pp. 5689-5700, Apr. 2023.
[28] J. Wang, H. Cai, S. You *et al*., "A framework for techno-economic assessment of demand-side power-to-heat solutions in low-temperature district heating," *International Journal of Electrical Power & Energy Systems*, vol. 122, p. 106096, Nov. 2020.
[29] Y. Li. (2017, Jan.). Deep reinforcement learning: an overview. [Online]. Available: https://arxiv.org/abs/1701.07274
[30] R. Lowe, Y. I. Wu, A. Tamar, *et al*., "Multi-agent actor-critic for mixed cooperative-competitive environments," *Advances in Neural Information Processing Systems*, vol. 30, pp. 6379-6390, Jan. 2017.
[31] T. P. Lillicrap. (2015, Feb.). Continuous control with deep reinforcement learning. [Online]. Available: https://arxiv.org/abs/1509.02971
[32] M. E. Baran and F. F. Wu, "Network reconfiguration in distribution systems for loss reduction and load balancing," *IEEE Transactions on Power Delivery*, vol. 4, no. 2, pp. 1401-1407, Apr. 1989.
[33] Australia Renewable Energy Agency. (2022, Nov.). Demonstration of three dynamic grid-side technologies. [Online]. Available: https://arena.gov.au/projects/demonstration-of-three-dynamic-grid-side-technologies/
[34] J. I. Nielsen, P. B. Nøgård, and C. Greisen. (2020, Jan.). Results from an urban living lab. [Online]. Available: https://www. energylabnordhavn. com/uploads/3/9/5/5/39555879/energylab_nordhavn_final_report_2020.pdf
[35] J. Wang. (2025, Jan.). System parameter. [Online]. Available: https://github.com/Jiawei37/DDPG_IEHS/blob/main/System%20parameters.pdf
[36] D. P. Kingma and J. Ba. (2014, Mar.). Adam: a method for stochastic optimization. [Online]. Available: https://arxiv.org/abs/1412.6980
[37] J. Schulman, F. Wolski, P. Dhariwal *et al*. (2017, Feb.). Proximal policy optimization algorithms. [Online]. Available: https://arxiv. org/abs/1707.06347
[38] R. D. Zimmerman, C. E. Murillo-Sanchez, and R. J. Thomas, "MATPOWER: steady-state operations, planning, and analysis tools for power systems research and education," *IEEE Transactions on Power Systems*, vol. 26, no. 1, pp. 12-19, Feb. 2011.

Hangyue Liu received the M.E. degree from the University of Sydney, Sydney, Australia, in 2017, and the Ph.D. degree in electrical engineering from The University of New South Wales, New South Wales, Australia, in 2023. He is currently working in the Sungrow Power Supply Co., Ltd., Hefei, China. His research interests include power system analysis, advanced power electronics control, and application of artificial intelligence in power systems.

Cuo Zhang received the B.E. (Hons.) degree in electrical (power) engineering from The University of Sydney, Sydney, Australia, in 2014, and the Ph. D. degree in electrical engineering from The University of New South Wales, New South Wales, Australia, in 2018. He is a Lecturer in power engineering at The University of Sydney. He is also a Discovery Early Career Researcher Award (DECRA) Fellow of Australian Research Council (ARC). His research interests include power system planning and operation, voltage stability and control, transactive energy, distributed energy resource, microgrid, and application of optimization theory and artificial intelligence in these areas.

Jiawei Wang received the M.Sc. and Ph.D. degrees in electrical engineering from Technical University of Denmark (DTU), Lyngby, Denmark, in 2016 and 2020, respectively. She is an Associate Professor with the National Frontiers Science Center for Industrial Intelligence and System Optimization, Northeastern University, Shenyang, China. Before that, she was an Assistant Professor with Northumbria University, Newcastle upon Tyne, U.K., a Postdoctoral Researcher with DTU, and a Visiting Postdoctoral Fellow with Imperial College London, London, U.K.. Her research interests include artificial intelligence and control theory applied to energy systems.

Ke Meng received the Ph.D. degree in electrical engineering from The University of Queensland, Brisbane, Australia, in 2009. He is currently the Technical Director of OSA Engineering Pty Ltd., Perth, Australia, and Senior Power System Engineer at ElectraNet Pty Ltd., Adelaide, Australia. Previously, he was a Senior Lecturer in energy systems at the School of Electrical Engineering and Telecommunications, The University of New South Wales, Sydney, Australia. His research interests include electric power system modeling, stability analysis, renewable energy system, and grid integration.

Zhao Yang Dong received the Ph. D. degree from The University of Sydney, Sydney, Australia, in 1999. He is currently a Chair Professor and the Head of Department of Electrical Engineering, and he is also Global STEM Professor and Director of JC STEM Lab of Future Energy Systems at City

University of Hong Kong, Hong Kong, China. His immediate roles were Singapore Power Group endowed Professor of Power Engineering, and Co-Director of SPG-NTU Joint Lab at Nanyang Technological University, Singapore. His previous roles include a SHARP Professor of energy systems, the inaugural Director of The University of New South Wales Digital Grid Future Institute and The University of New South Wales, Sydney, Australia, the Director of the ARC Research Hub for Integrated Energy Storage Systems, the Head of the School of Electrical and Information Engineering, The University of Sydney, and the Ausgrid Chair Professor and the Director of the Ausgrid Centre for Intelligent Electricity Networks. His research interests include power system planning and stability, smart grid, smart city, renewable energy system, and energy market.