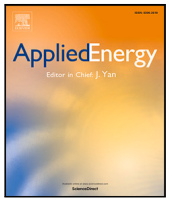
Applied Energy 336 (2023) 120826

Contents lists available at ScienceDirect

Applied Energy

journal homepage: www.elsevier.com/locate/apenergy



Hierarchical multi-agent reinforcement learning for repair crews dispatch control towards multi-energy microgrid resilience

Dawei Qiu [a], Yi Wang [a,*], Tingqi Zhang [b,c], Mingyang Sun [c], Goran Strbac [a]

[a] Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, UK
[b] Electric Power Research Institute, State Grid Liaoning Electric Power Company Ltd., Shenyang, 110000, China
[c] Department of Control Science and Engineering, Zhejiang University, Hangzhou 310027, China

ARTICLE INFO

Keywords:
Repair crews
Multi-energy microgrid
Resilience
Power-gas-transportation network
Hierarchical multi-agent reinforcement learning

ABSTRACT

Extreme events are greatly impacting the normal operations of microgrids, which can lead to severe outages and affect the continuous supply of energy to customers, incurring substantial restoration costs. Repair crews (RCs) are regarded as crucial resources to provide system resilience owing to their mobility and flexibility characteristics in handling both transportation and energy systems. Nevertheless, effectively coordinating the dispatch of RCs towards system resilience is a complex decision-making problem, especially in the context of a multi-energy microgrid (MEMG) with enormous dynamics and uncertainties. To this end, this paper formulates the dispatch problem of RCs in a coupled transportation and power-gas network as a decentralized partially observable Markov decision process (Dec-POMDP). To solve this Dec-POMDP, a hierarchical multi-agent reinforcement learning (MARL) algorithm is proposed by featuring a two-level framework, where the high-level action is used for switching decision-making between transportation and power-gas networks, and the lower-level action constructed via the multi-agent proximal policy optimization (MAPPO) algorithm is used to compute the routing and repairing decisions of RCs in the transportation and power-gas networks, respectively. The proposed algorithm also introduces an abstracted critic network by integrating the load restoration status, which captures the system dynamics and stabilizes the training performance with privacy protection. Extensive case studies are evaluated on a coupled 6-bus power and 6-bus gas network integrated with a 9-node 12-edge transportation network. The proposed algorithm outperforms the conventional MARL algorithms in terms of policy quality, learning stability, and computational performance. Furthermore, the dispatch strategies of RCs are analyzed and their corresponding benefits for load restoration are also evaluated. Finally, the scalability of the proposed method is also investigated for a larger 33-bus power and 15-bus gas network integrated with an 18-node 27-edge transportation network.

1. Introduction

Since the increasing number of high-impact and low-probability (HILP) events, power network resilience has received much research interest from power industries and researchers [1]. HILP events include both natural and man-made disasters (e.g., hurricanes, earthquakes, cyber-attacks, etc.), which can cause a fatal impact on power networks [2]. Much research has focused on the development of effective planning and operation strategies for resilience enhancement under the concept of microgrids (MGs) [3]. Specifically, using locally available distributed energy resources (DERs) (e.g., renewable energy resources (RESs), energy storage systems (ESSs), and diesel generators (DGs)) to restore essential loads can effectively improve resilience and reduce economic losses after a serious outage [4]. However, it is worth noting that an extreme event may also influence the natural gas network and interrupt gas supplies to customers. Thus, growing attention has been paid to the interdependence between power networks and natural gas networks in deciding the restoration strategy [5].

In addition to various DERs, repair crews (RCs) are crucial resources for energy system outage management against natural disasters [6]. It is desired that RCs repair damaged components (e.g., power plants and network lines) in an optimal order; thus, the outage time can be reduced, the operating system can be recovered, and more essential loads can be restored in a short time. In general, the RC dispatching problem can be divided into two sub-problems: routing and repairing [7], which is a challenging task for efficient and stable system operations considering the complexity of capturing the coupled transportation, power, and gas networks. On the other hand, decentralization and digitization are rapidly transforming the energy sector and disrupting

* Corresponding author.
E-mail address: yi.wang18@imperial.ac.uk (Y. Wang).

https://doi.org/10.1016/j.apenergy.2023.120826
Received 12 July 2022; Received in revised form 6 November 2022; Accepted 7 February 2023
Available online 14 February 2023

Nomenclature

A. Indexes and Sets

$t \in T$	Index and set of time steps
$i \in I$	Index and set of repair crews (RCs)
$d \in ED$	Index and set of electric demand (ED)
$d \in GD$	Index and set of gas demand (GD)
$g \in DG$	Index and set of diesel generators (DGs)
$g \in GG$	Index and set of gas-fired generators (GGs)
$g \in PV$	Index and set of PV generators
$g \in GW$	Index and set of gas wells (GWs)
$b \in EB$	Index and set of electric buses (EBs)
$b \in GB$	Index and set of gas buses (GBs)
$n \in N$	Index and set of transportation nodes
$l \in PL$	Index and set of distribution lines in power network (PLs)
$l \in GL$	Index and set of pipelines in gas network (GLs)
$r \in R$	Index and set of transportation roads

B. Parameters

Δt	Time resolution (1 h)
c_d^{pow}	Load shedding cost of ED d (£/kWh)
c_d^{gas}	Load shedding cost of GD d (£/S m^3)
$P_{g,t}^{pv}$	Active power of PV g at time step t (kW)
$P_{d,t}^{ed}$	Active power of ED d at time step t (kW)
$Q_{d,t}^{ed}$	Reactive power of ED d at time step t (kVAR)
$G_{d,t}^{gd}$	Gas level of GD d at time step t (S m^3)
$\overline{P}_g^{dg}$	Maximum active power of DG g (kW)
$\overline{Q}_g^{dg}$	Maximum reactive power of DG g (kVAR)
$\underline{P}_g^{dg}$	Minimum active power of DG g (kW)
$\underline{Q}_g^{dg}$	Minimum reactive power of DG g (kVAR)
$\overline{P}_g^{gg}$	Maximum power of GG g (kW)
$\underline{P}_g^{gg}$	Minimum power of GG g (kW)
b_g^{gg}	Coefficient for gas consumption of GG g (S m^3/kW)
$\overline{G}_g^{gw}$	Maximum gas output of GW g (S m^3)
$\underline{G}_g^{gw}$	Minimum gas output of GW g (S m^3)
$\overline{V}$	Maximum permissible voltage (p.u.)
$\underline{V}$	Minimum permissible voltage (p.u.)
B_l	Susceptance of power line l (p.u.)
G_l	Conductance of power line l (p.u.)
$\overline{S}_l$	Capacity limit of power line l (kVA)
$\overline{G}_l$	Capacity limit of gas pipeline l (S m^3)
λ_l	Compression factor of compressor on gas pipeline l
η_l	Parameter for gas flow and pressure relationship on gas pipeline l (S m^3/(h bar^2))
RT_i^{rc}	Required time period for RC i to repair damaged components (h)
RS_i^{rc}	Resource capacity of RC i to repair damaged components (unit)

C. Variables

$P_{d,t}^{ls}$	Load shedding of ED d at time step t (kW)
$G_{d,t}^{ls}$	Load shedding of GD d at time step t (S m^3)
$P_{g,t}^{dg}$	Active power generation of DG g at time step t (kW)
$Q_{g,t}^{dg}$	Reactive power generation of DG g at time step t (kVAR)
$P_{g,t}^{gg}$	Power generation of GG g at time step t (kW)
$Q_{g,t}^{gg}$	Gas consumption of GG g at time step t (S m^3)
$G_{g,t}^{gw}$	Gas output of GW g at time step t (S m^3)
$V_{b,t}$	Voltage of electric bus b at time step t (p.u.)
$\delta_{b,t}$	Voltage angle of electric bus b at time step t (°)
$P_{b,t}^{ex}$	Active power exchange between electric bus b and other buses at time step t (kW)
$Q_{b,t}^{ex}$	Reactive power exchange between electric bus b and other buses at time step t (kVAR)
$P_{l,t}$	Active power of power line l at step t (kW)
$Q_{l,t}$	Reactive power of power line l at step t (kVAR)
$P_{l,t}^{lo}$	Active power loss through power line l at time step t (kW)
$Q_{bp,t}^{lo}$	Reactive power loss through power line l at time step t (kVAR)
$G_{l,t}$	Gas flow of gas pipeline l at time step t (S m^3)
$\rho_{b,t}$	Gas pressure of nodal b at time step t (bar)

the traditional top-down philosophy of power systems [8], which makes it a necessity to develop an effective distributed control algorithm for those small-size decentralized RCs to exploit their flexibility in resilience enhancement. To the best of the authors' knowledge, the integrated routing and repairing characteristics of RCs in a decentralized framework for the coupled power-gas-transportation network have not been investigated under the topic of resilience enhancement. As such, this paper tries to propose a real-time and automatic control strategy for a collaborative RCs dispatching problem towards multi-energy system resilience enhancement in the context of a coupled power-gas-transportation network.

1.1. Literature review on RCs for resilience enhancement

Model-based optimization methods have been widely developed to formulate the routing and repairing characteristics of RCs on providing resilience, e.g., load restoration problems [9]. For instance, in [10], a distribution system restoration model involving RC dispatch and networked MGs is suggested in response to multiple line outages caused by natural disasters. In [11], the routing and repairing characteristics of RCs are studied via an integrated mixed integer-linear programming (MILP) model for the resilient service restoration of unbalanced distribution systems. However, uncertainties are not captured in the above two papers. In [12], a two-stage MILP model is proposed for the resilience-driven dispatching problem of RCs towards load restoration. Uncertainties associated with system demand and repair time are handled via stochastic programming (SP). In [13], a co-dispatch formulation of RCs and mobile emergency generators (MEGs) is proposed to enhance both the survivability and restoration capability of distribution systems. Uncertain network damage is modeled via the SP as well. In [14], a hybrid stochastic-robust optimization approach is developed to deal with various uncertainties (e.g., travel and repair time, demand, and renewable energies) in the co-optimization problem of RCs dispatch and distribution network reconfiguration. However,

it is worth noting that the above research only focuses on power networks but ignores the integration of power networks with other sectors, e.g., natural gas networks.

Recently, the increasing penetration of gas-fired generators and the use of electricity-driven gas compressors have increased the interdependence between power and natural gas networks, which raises the importance of considering the resilience enhancement of combined networks [15]. Only a few studies are focusing on the coordination effects of RCs and coupled power-gas networks. In [7], a combined RC dispatch problem is proposed to improve the resilience of interdependent power and natural gas systems in a post-disaster repair. However, the proposed model does not consider any system uncertainty. In [16], a sequential restoration model is developed to determine the repair sequence of damaged components for resilience maximization, where the restoration characteristics in terms of repair modes, repair time, and recovery costs are considered. However, the detailed routing behaviors of RCs and the associated uncertainties are not modeled in this paper. In [17], a service-based optimal energy flow model is proposed to minimize the consequences caused by windstorms through the coordination of RCs and multi-energy systems. The uncertain locations of damaged components are captured via the Monte-Carlo simulation technique.

It can be concluded that extensive efforts have been made to study the RC routing and repairing problems for the resilience enhancement of the coupled power-gas networks. However, the limitations of the above research cannot be erased and are summarized hereafter. Firstly, the routing characteristics of RCs are not settled in a detailed transportation environment; thus, road congestion impact is not considered. Secondly, uncertainties are handled via SP or robust optimization (RO) approaches, which may only be able to capture a small number of representative scenarios or lead to very conservative optimization results. Meanwhile, SP approaches can be time-consuming, especially when a large number of scenarios are involved; hence, they are not capable of providing timely services for resilient multi-energy systems. Thirdly, the model-based optimization problems assume that RCs require complete knowledge of the experiment environment, e.g., power network, gas network, transportation status, accurate uncertain parameters, etc. However, such assumptions are normally impractical considering the highly stochastic and dynamic real-world environment.

1.2. Literature review on state-of-the-art RL

Given the shortcomings of model-based optimization approaches, *reinforcement learning* (RL) [18] is regarded as a model-free approach to studying the sequential and dynamic decision-making problems of agents who can gradually learn the optimal control decisions by utilizing experiences gained from their repeated interactions with the environment, without a *prior* knowledge. In addition, RL as an online learning method can make efficient use of increasing data, thereby capturing the system uncertainties and adapting to various state dynamics. Finally, once the RL algorithm is well trained, its policy can be delivered to the online test set on timescales of milliseconds without requiring any identification. Therefore, RL is claimed as an efficient tool for real-time automatic control applications.

Previous research has successfully applied various RL algorithms to energy system applications [19,20], but few of them have focused on MG resilience issues. Overall, RL can be classified into two categories, of which the first one focuses on the problem of a single entity, employing *single-agent reinforcement learning* (SARL) algorithms. In [21], a Q-learning (QL) algorithm is used to generate the sequential order of repairing damaged components and update the network topology to obtain the largest amount of power supply on a given network topology. In [22], a twin delayed deep deterministic policy gradient (TD3) algorithm is applied to train the control decisions of mobile energy storage systems (MESSs) designated destination and charging/discharging behaviors, as well as the generation output, towards the MG resilience. In [23], a model-free optimization framework based on the proximal policy optimization (PPO) algorithm is proposed to determine the optimal rescheduling strategy of distributed generators and controllable loads to improve the resilience of a distribution system. In [24], a deep deterministic policy gradient (DDPG) algorithm is proposed to optimize the power dispatches of generators and the charging/discharging of ESSs. In [25], an advantage actor–critic (A2C) algorithm is proposed for grid hardening decisions to improve the long-term distribution system resilience. Nevertheless, the above research has successfully applied various RL algorithms to power system resilience, none of which considers the RC applications. To the best of the authors' knowledge, there is only one paper [26] that has applied the DQN algorithm to assess the effectiveness of a repair crew allocation strategy and optimize the strategy after an extreme event to improve the infrastructure network resilience. Besides the conventional DRL algorithms, transfer learning is a kind of technique learning the control policy from a certain level of knowledge that can help accelerate the convergence speed, which has been applied for the real-time energy management of plug-in hybrid vehicles [27,28].

The second category focuses on the problem of a multi-agent setup, employing *multi-agent reinforcement learning* (MARL) algorithms. In [29], a double-agent RL model based on DQN and DDPG is proposed to optimize the discrete load shedding control and the continuous power generation in an islanded MG, with the objective of continuous power supply to the distribution system. However, DQN and DDPG are two separate RL algorithms that may bring out the inconsistency of a multi-agent setup. Furthermore, the load agents and generator agents train their control policies independently, which may cause the instability of the training performance, since the policies are interactional in a multi-agent setup. To this end, a multi-agent parameterized double DQN method is proposed in [30] that introduces an abstracted critic network incorporating the load restoration conditions, thereby capturing the system dynamics and stabilizing the training performance. Furthermore, instead of using two separate RL algorithms, this paper computes a hybrid policy for both (discrete) routing and (continuous) scheduling decisions of multiple MESS agents in a coupled power-transportation network, to improve MG load restoration. Apart from controlling the energy resources (e.g., load, generators, and storage), MARL algorithms have been applied to network reconfiguration problems for system resilience. In [31], a multi-agent soft actor–critic (MASAC) algorithm is proposed to control the operating switches of the MG network system to improve load restoration. In [32], a multi-agent-based hybrid soft actor–critic (HSAC) algorithm is developed for the siting and sizing of shunt reactive power compensators to enhance system voltage resilience.

The above literature has successfully applied various SARL and MARL algorithms to many promising MG resilience problems. Nevertheless, both two categories neglect the application of RCs to multi-energy microgrid (MEMG) resilience enhancement in the context of a coupled power-gas-transportation network. However, RC as a mobile and flexible resource plays a vital role in MEMG resilience enhancement. On the one hand, RCs can move to any location in the transportation network where the energy components are damaged. On the other hand, RCs can make use of their efficient repairing ability to resume grid operation and ensure a reliable system. However, making efficient routing and repairing decisions in both transportation and power-gas networks is not straightforward, since the routing and repairing decisions of RCs are mutually exclusive, i.e., RCs cannot repair during the trip in the transportation network, and also cannot drive when connected to the grid in a power-gas network.

1.3. Research gaps and contributions

Although the above studies have demonstrated the valid applications of both model-based optimization and model-free MARL methods to the RC dispatch problem, the following fundamental research gaps still remain:

(1) Previous work [7,10–17] employs model-based optimization approaches to solve the resilience-oriented dispatch problem of RCs, which require complete knowledge of the experiment environment (e.g., power network, gas network, transportation status, accurate uncertain parameters, etc.) and can be time-consuming. Both of these characteristics are impractical, especially when HILP events are taken into account.
(2) Previous work applies SARL [22–25] and MARL [29–32] methods for various DER scheduling and management problems; nevertheless, there is no research focusing on the application of RL algorithms to resilience-oriented RC dispatch problems in a highly stochastic and dynamic environment.
(3) Previous work employs a variety of MARL algorithms, e.g., MADQN [29], MADDGN [30], and MADDPG [29], MASAC [31,32], to learn the operating strategies towards resilience enhancement. However, the investigated RC dispatch problem involves both routing and repairing decisions, which correspond to the transportation and power networks, respectively. Since the routing and repairing decisions of RCs are mutually exclusive, the above MARL algorithms fail to fully exploit the flexibility of these two decisions.

To fill in the research gaps discussed above, this paper introduces a hierarchical-based MARL algorithm that first learns a high-level (HL) policy for selecting either making routing actions in a transportation network or making repairing actions in a power-gas network; and then learns a lower-level (LL) policy for making specific routing and repairing actions. More specifically, the novel contributions of this paper are described below:

(1) A *decentralized partially observable Markov decision process* (Dec-POMDP) [18] is proposed to formulate the resilience enhancement problem of a coupled power-gas-transportation network for the coordination effect of multiple RCs on repairing damaged components and reducing load shedding after extreme events. The decentralized formulation can significantly reduce the computational burden, avoid potential privacy issues, and ensure reasonable RC dispatching results.
(2) Transportation, power, and gas networks are modeled as the simulation environment of the proposed Dec-POMDP for RCs' realistic decision-making process. In detail, a traffic network model is proposed to capture the impact of traffic time and congestion, while a linearized network model is employed to capture the coupling operation of power and gas networks. RCs make the route decisions and repair schedules in this complex power-gas-transportation network. In such a case, both the mobility and flexibility of RCs can be exploited efficiently.
(3) A novel MARL method, namely HMAPPO, is proposed to efficiently solve this Dec-POMDP by proposing a hierarchical learning architecture [33] to select either making a routing or a repairing decision; a multi-agent proximal policy optimization (MAPPO) algorithm [34] with parameter-sharing (PS) technique [35] to stabilize the training performance and enhance the learning scalability; and a collective restoration index [30] to represent the system dynamics with privacy protection.
(4) Extensive case studies on a real-world dataset have been developed to evaluate the superior performance of the proposed HMAPPO over the conventional MARL algorithms in terms of policy quality, learning stability, and computational time. A generalized and real-time automatic routing and repairing policy is produced and can be adapted to various power-gas-transportation network uncertainties associated with traffic volumes, demand profiles, and renewable energies. Finally, the proposed method demonstrates its scalability to different RC and network sizes.

1.4. Paper organization

The rest of this paper is organized as follows. Section 2 describes the studied problem and presents the mathematical formulations of RC routing and repairing models in the coupled power-gas-transportation network. Section 3 formulates the RC-related resilience enhancement problem as a Dec-POMDP. Section 4 provides the detailed algorithm of the proposed HMAPPO that can efficiently solve the Dec-POMDP. The experimental setup and the case studies are presented in Sections 5 and 6, respectively. Finally, Section 7 discusses and concludes this work as well as the future extensions.

2. Problem formulations of RC models in the coupled power-gas-transportation network

We focus on the joint routing and repairing problem of multiple RCs for MEMG resilience enhancement (e.g., load restoration) in a coupled power-gas-transportation network, as illustrated in Fig. 1, which involves three modules: the communication layer, the transportation network, and the coupled power-gas network. In more detail, there is a central monitor that can communicate with each RC and broadcast to individuals the system outage information (e.g., damaged component, status, and location) via the communication layer, in order to help RCs make more informative decisions. However, unlike the previous research [7] that RCs follow the command of the central monitor (e.g., dispatch center), these privacy-preserving RCs are operated in a decentralized manner in anticipation of a future trend towards resilient distribution networks [36]. As a result, the central monitor is unable to acquire the local information of RCs as long as each RC's local information is kept private. Then, these RCs can move between different transportation nodes and choose to repair certain damaged components. Inside the power network, DERs are appropriately deployed on certain buses, including electric demand (ED), diesel generators (DGs), and photovoltaics (PVs). Inside the gas network, gas demand (GD) and gas wells (GWs) are appropriately deployed on certain buses. The power network and gas network are coupled through gas-fired generators (GGs).

Specifically, the potential destinations of RCs include their initial depots and the locations of damaged components (e.g., power plants, distribution lines, and gas pipelines). When an extreme event happens, RCs are sent out (gray module in Fig. 1) to sequentially repair the damaged components, which constitutes two kinds of alternative decisions: (1) routing to destinations in the transportation network and (2) repairing the targeted components in the power-gas network. On one hand, for each journey in the transportation network (blue module in Fig. 1), after observing the real-time transportation information of map location and traffic volumes, RCs with a smart routing algorithm can optimally manage the moving directions in the transportation network. On the other hand, for each operation time reaching a certain damaged component in the power-gas network (green and orange modules in Fig. 1), after observing the real-time power or gas information of nearby demand and RESs, as well as the required resources and time horizon to repair this component, RCs with a smart repairing algorithm can optimally choose to repair the component or not. However, these two decisions are mutually influenced since making an efficient route plan can save more time for RCs to fully exploit their carried resources in resilience enhancement, while making an efficient repair decision can allow RCs to have more time to exploit their transportation mobility.

Once the parking locations of RCs are determined, the coupled power-gas network equips a microgrid central controller (MGCC) that can optimally manage the power and gas schedules of each controllable component by maximizing the overall system load restoration. Additionally, the coupled system is fully modeled by a linearized network model capturing all the network constraints and technical constraints related to stability properties, which ensures accurate optimization results and secure system operations. In this section, we present the details of the proposed mathematical models, including the RCs' routing and repairing, the transportation network, and the linearized network model for the coupled power-gas network.

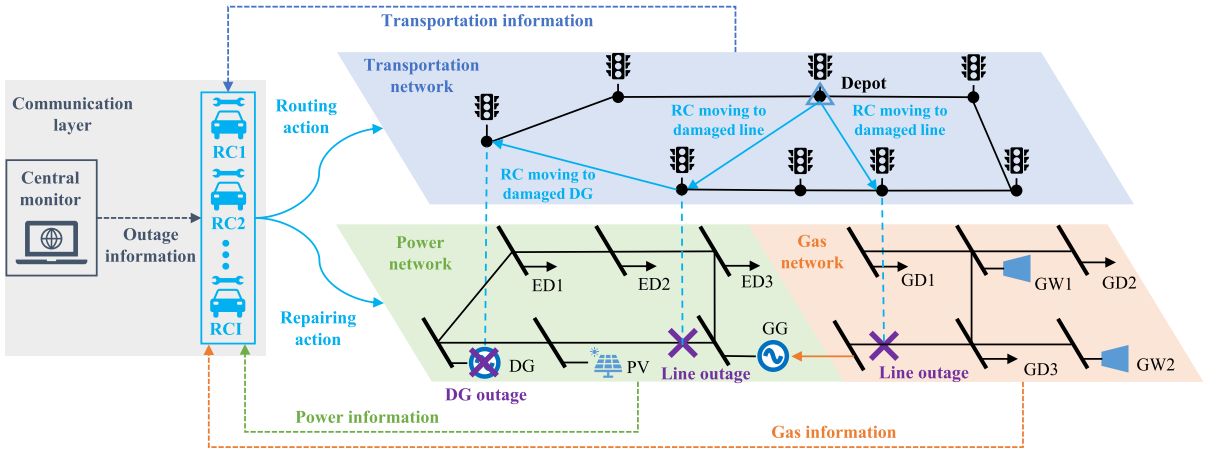


Fig. 1. The scheme of RCs routing and repairing process for multi-energy system resilience enhancement of a coupled power-gas-transportation network.

2.1. RCs routing and repairing

Let N be the set of original transportation nodes, where subsets N^{dm} are the set of candidate nodes appropriately selected for RC routing, i.e., damaged components [6]. I represent the set of RCs. The routing behavior of RCs in the set of candidate transportation nodes N^{dm} can be formulated by (1)–(2), which is sufficient for the incorporation of the routing problem into an optimization problem [6]. In more detail, constraint (1) demonstrates that RC i can only be connected with one candidate node at one time step t, where binary variable $\alpha^{rc}_{i,m,t}$ indicates whether RC i is connected with the transportation node m at time step t (if $\alpha^{rc}_{i,m,t}=1$) or not (if $\alpha^{rc}_{i,m,t}=0$). Constraint (2) ensures that RC i routes reasonably between different nodes, where T^{trl}_{mn} refers to the time period required to route from node m to node n. Note that the traveling time T^{trl}_{mn} can vary with real-time road congestion, which is discussed in the next subsection.

$$\sum_{m\in N^{dm}_i} \alpha^{rc}_{i,m,t} \le 1,\ \forall i\in I,\ \forall t\in T \tag{1}$$

$$\sum_{\tau=t}^{\min(t+T^{trl}_{mn},T)} \alpha^{rc}_{i,n,\tau} \le (1-\alpha^{rc}_{i,m,t})\cdot \min(T^{trl}_{mn}, T-t),\ \forall i\in I,\ \forall m\in N^{dm}_i,$$
$$\forall n\in N^{dm}_i\backslash\{m\},\ \forall t\in T \tag{2}$$

Depending on the routing behaviors of RCs, the repair plan of RC i is formulated in constraints (3) and (4). More specifically, the binary variable $z^{rc}_{i,m,t}$ is 1 if the damaged component m is repaired by RC i at time step t, 0 otherwise, and $RT^{rc}_{i,m}$ corresponds to the time period required to repair component m [6]. Constraint (4) ensures that the damaged component m remains intact if it has been repaired by RC i. Constraint (5) ensures that RC i's resource capacity S^{rc}_i is sufficient for its repair tasks, where $RS_{i,m}$ represents the resources required by RC i to repair the damaged component m.

$$z^{rc}_{i,m,t} \le \Big(\sum_{\tau=1}^{t} \alpha^{rc}_{i,m,\tau}\Big)/RT^{rc}_{i,m}, \forall i\in I, \forall m\in N^{dm}_i, \forall t\in T \tag{3}$$

$$z^{rc}_{i,m,t} \le z^{rc}_{i,m,t+1},\ \forall i\in I,\ \forall m\in N^{dm}_i,\ \forall t\le T-1 \tag{4}$$

$$\sum_{m\in N_{rc}} RS_{i,m}\cdot z^{rc}_{i,m,T} \le S^{rc}_i,\ \forall i\in I \tag{5}$$

2.2. Transportation network modeling

In the studied problem, RCs have different types of candidate nodes (i.e., damaged lines and resources), which can formulate their own transportation networks constituted by N^{dm}. Here, we illustrate a simple example in Fig. 2, of which the RC i's transportation network N^{dm}_i can be formulated as the closed-loop N1–N2–N5–N4.

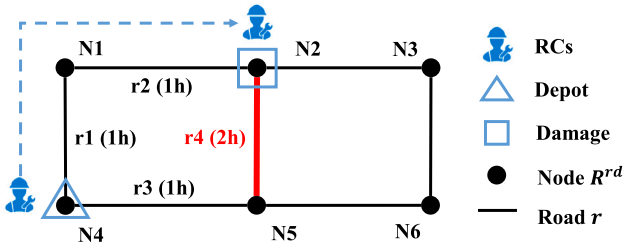


Fig. 2. Routing characteristics of RCs in a transportation network.

Specifically, RC i performs routing behaviors on its own transportation network topology N^{dm}_i, which are influenced by the node locations $R^{rd}_{\{1,2,4,5\}} \in N^{dm}_i$ and the travel time $T^{trl}_{r,t}$ between any two candidate nodes on road r at time step t. It should be noted that the node locations R^{rd} are fixed in relation to the network topology, whereas the travel time $T^{trl}_{r,t}$ is influenced by real-time traffic volumes, which can be estimated using [37] as follows:

$$T^{trl}_{r,t} = \tilde{T}^{trl}_r\Big[1+\alpha^{rd}\Big(\frac{V^{rd}_{r,t}}{C_r}\Big)^{\beta^{rd}}\Big],\ \forall r\in R,\ \forall t\in T \tag{6}$$

where $\tilde{T}^{trl}_r$ is the traveling time of free prevailing driving that mainly depends on the road length. C_r, α^{rd}, β^{rd} and $V^{trl}_{r,t}$ correspond to the capacity, the retardation coefficients and the real-time traffic volume of road r at time step t, respectively. Sets R and T indicate the road set of the transportation network and the time set of the daily horizon, respectively. This function can describe the relationship between traveling time and traffic volume but also reflect road impedance from the characteristics of traffic flow itself. Note that it is very important to consider the road congestion impact for realistic routing behaviors; nevertheless, most existing literature on MPS routing problems [6,10,11] ignores this factor.

In order to clearly illustrate the routing characteristics simulated in this paper, we prepare an example in this subsection. It can be observed from Fig. 2 that the depot location of RC i is N4 and the location of the damaged line is N2. Within the network topology N^{dm}_i, there are two available routes for RC i to commute (i.e., i.e., $r_1\to r_2$ and $r_3\to r_4$), but serious congestion occurs on road r_4 (the red road in Fig. 2), resulting in a much longer travel time (3 h) on route $r_3\to r_4$. In this context, the RC will choose $r_1\to r_2$ as its commuting route rather than $r_3\to r_4$ due to the less 1 h traveling time. Finally, the RC is expected to arrive at its repairing destination, which is represented by the transportation node set $\{R^{rd}_2\}$.

2.3. Coupled power-gas network modeling

To incorporate appropriate RC dispatching behaviors into a coupled power-gas network, this subsection introduces a linearized model for network operations, capturing real-time statuses of damaged components. Specifically, after RC i makes real-time repair decisions, the network operator can solve the following linearized network model for each time step t, which includes both the power network and the natural gas network.

$$\min_{\Xi^{opf}} \mathbb{E}\{\sum_{d\in EL} c_{d,t}^{pow} P_{d,t}^{ls} + \sum_{d\in GL} c_{d,t}^{gas} G_{d,t}^{ls}\} \tag{7}$$

where

$$\Xi^{opf} = \{P_{g,t}^{dg}, Q_{g,t}^{dg}, P_{d,t}^{ls}, P_{b,t}^{ex}, Q_{b,t}^{ex}, P_{bp,t}, Q_{bp,t}, V_{b,t}^2, \delta_{bp,t}, G_{g,t}^{gg}, G_{g,t}^{gw}, G_{d,t}^{ls}, G_{l,t}, \rho_{b,t}\}, \tag{8}$$

where the objective function (7) towards the expectation of load shedding minimization includes two terms: (1) shedding cost of electric demand in the power network; (2) shedding cost of gas demand in the gas network. The randomness of system uncertainty parameters (e.g., demand, PV power generation, and damaged components) and corresponding stochastic decision variables (e.g., power flows, gas flows, generation outputs, gas well outputs, voltage angles, and gas pressure) is taken into account by the expectation operator $\mathbb{E}\{\cdot\}$.

The optimization problem of power network is subject to the active power balance at exchange bus b presented in (9), while the reactive power balance corresponds to (10). The sets B_{ed}, B_{dg}, B_{pv}, and B_{gg} correspond to the nodal ED, DG, PV, and GG located at bus b, respectively. Classical equations pertaining to power flow problems are linearized and shown in (11)–(12). Eqs. (13)–(14) correspond to the power network model, in which $P_{bp,t}$ and $Q_{bp,t}$ are the active and reactive power passing the line $b-p$, respectively, and $P_{bp,t}^{lo}$ and $Q_{bp,t}^{lo}$ are related to the losses from active and reactive power linearized by loss factors according to [38]. Constraints (15) and (16) are related to the operation characteristics of voltage limits at bus b and thermal capacities for line l, respectively. Finally, constraints (17)–(18) and (19)–(20) show the active and reactive power limits of DG g and GG g, respectively. Binary variable $z_{bp,t}^{pl}$, $z_{g,t}^{dg}$, and $z_{g,t}^{gg}$ correspond to the status of power line $b-p$, DG g, and GG g, respectively (1 if intact, 0 if not), depending on the RC repairing behaviors shown in Eq. (3).

$$\sum_{g\in B_{dg}} P_{g,t}^{dg} + \sum_{g\in B_{ed}} P_{d,t}^{ls} + \sum_{g\in B_{pv}} P_{g,t}^{pv} + \sum_{g\in B_{gg}} P_{g,t}^{gg} = P_{b,t}^{ex} + \sum_{d\in B_{ed}} P_{d,t}^{ed}, \quad \forall b\in EB,\ \forall t\in T \tag{9}$$

$$\sum_{g\in B_{dg}} Q_{g,t}^{dg} + \sum_{g\in B_{gg}} Q_{g,t}^{gg} = Q_{b,t}^{ex} + \sum_{d\in B_{ed}} Q_{d,t}^{ed},\ \forall b\in EB,\ \forall t\in T \tag{10}$$

$$P_{b,t}^{ex} = \sum_{(b,p)\in L} P_{bp,t} + (\sum_{p\in B} G_{bp})V_{b,t}^2,\ \forall b\in EB,\ \forall t\in T \tag{11}$$

$$Q_{b,t}^{ex} = \sum_{(b,p)\in L} Q_{bp,t} - (\sum_{p\in B} B_{bp})V_{b,t}^2,\ \forall b\in EB,\ \forall t\in T \tag{12}$$

$$P_{bp,t} = G_{bp}(V_{b,t}^2 - V_{p,t}^2)/2 - B_{bp}\delta_{bp,t} + P_{bp,t}^{lo},\ \forall bp=l\in PL,\ \forall t\in T \tag{13}$$

$$Q_{bp,t} = -B_{bp}(V_{b,t}^2 - V_{p,t}^2)/2 - G_{bp}\delta_{bp,t} + Q_{bp,t}^{lo},\ \forall bp=l\in PL,\ \forall t\in T \tag{14}$$

$$\underline{V}^2 \le V_{b,t}^2 \le \overline{V}^2,\ \forall b\in EB,\ \forall t\in T \tag{15}$$

$$P_{bp,t}^2 + Q_{bp,t}^2 \le z_{bp,t}^{pl}\overline{S}_{bp},\ \forall bp=l\in PL,\ \forall t\in T \tag{16}$$

$$z_{g,t}^{dg}\underline{P}_g^{dg} \le P_{g,t}^{dg} \le z_{g,t}^{dg}\overline{P}_g^{dg},\ \forall g\in DG,\ \forall t\in T \tag{17}$$

$$z_{g,t}^{dg}\underline{Q}_g^{dg} \le Q_{g,t}^{dg} \le z_{g,t}^{dg}\overline{Q}_g^{dg},\ \forall g\in DG,\ \forall t\in T \tag{18}$$

$$z_{g,t}^{gg}\underline{P}_g^{gg} \le P_{g,t}^{gg} \le z_{g,t}^{gg}\overline{P}_g^{gg},\ \forall g\in GG,\ \forall t\in T \tag{19}$$

$$z_{g,t}^{gg}\underline{Q}_g^{gg} \le Q_{g,t}^{gg} \le z_{g,t}^{gg}\overline{Q}_g^{gg},\ \forall g\in GG,\ \forall t\in T \tag{20}$$

Regarding the natural gas network, a steady state natural gas operation is assumed in this paper [7]. Specifically, the gas system operation constraints include gas supply in (21), nodal gas balance in (22), and pipeline constraints in (23)–(26). Specifically, the capacity of GW is constrained by (21), while the nodal gas flow balance is shown in (22). The sets B_{gd}, B_{gw}, and B_{gg} correspond to the nodal GD, GW, and GG located at bus b, respectively. Pipelines without compressors are denoted as inactive pipelines belonging to GL^{ina}, while those with compressors are active pipelines belonging to GL^{act}. The nodal gas pressure $\rho_{b,t}$ for a compressor with the gas flow from b to p in GL^{act} is constrained by (23), where λ_l indicates the compressor's compression factor at pipeline l. Constraint (24) ensures that gas pressure at each node stays within a preset range. For an inactive gas pipeline with a gas flow from b to p in GL^{ina}, the relationship between gas flow and nodal gas pressure is represented by (25), where η_l represents the relationship between gas flow and pressure based on Weymouth equation. Furthermore, as expressed in (26), the gas flow is limited by pipeline capacity, where the binary variable $z_{l,t}^{gl}$ corresponds to the status of the gas pipeline l (1 if intact, 0 if not), depending on the RC repairing behavior as shown in Eq. (3). A piece-wise linearization technique from [39] is adopted to linearize the inactive pipeline constraint (25). Finally, Eq. (27) expresses the energy conversion between power generation P_g^{gg} and gas consumption G_g^{gg} of GG g, where b_g^{gg} represents the GG coefficient in Sm^3/kW and G_g^{gg} can be also regarded as the GD $G_{d,t}^{gd}$ of load l located at bus B_{gg} in the gas system.

$$\underline{G}_g^{gw} \le G_{g,t}^{gw} \le \overline{G}_g^{gw},\ \forall g\in GW,\ \forall t\in T \tag{21}$$

$$\sum_{g\in B_{gw}} G_{g,t}^{gw} + \sum_{d\in B_{gd}} G_{d,t}^{ls} = \sum_{d\in B_{gd}} G_{d,t}^{gd} + \sum_{pb\in GL} G_{pb,t} - \sum_{bp\in GL} G_{bp,t},\ \forall b\in GB,\ \forall t\in T \tag{22}$$

$$\rho_{b,t} \le \rho_{p,t} \le \lambda_l\rho_{b,t},\ \forall l\in GL^{act},\ \forall t\in T \tag{23}$$

$$\underline{\rho}_b \le \rho_{b,t} \le \overline{\rho}_b,\ \forall b\in GB,\ \forall t\in T \tag{24}$$

$$G_{l,t}^2 - \eta_l(\rho_{b,t}^2 - \rho_{p,t}^2) = 0,\ \forall l\in GL^{ina},\ \forall t\in T \tag{25}$$

$$0 \le G_{l,t} \le z_{l,t}^{gl}\overline{G}_l,\ \forall l\in GL,\ \forall t\in T \tag{26}$$

$$P_{g,t}^{gg} = G_{g,t}^{gg}/b_g^{gg} = G_{d,t}^{gd},\ \forall g\in GG,\ \forall t\in T,\ \forall d\in B_{gg} \tag{27}$$

2.4. Problem challenges and solutions

Solving the above coordinated dispatch problem of RCs within a power-gas-transportation network (Sections 2.2 and 2.3) is very challenging. First, the examined RCs become clueless if the technical parameters and mathematical models of the coupled networks are unknown due to the privacy concerns in the smart grid concept. Similarly, it is difficult to find the optimal solutions when utilities (e.g., MGCC) do not own RCs since their operational constraints are not integrated into the coupled network models. Second, the high-impact nature of extreme events necessitates capturing various uncertainties; nevertheless, it is impractical to obtain accurate probability distributions of uncertainties. Third, solving a time-coupled optimization problem may take a long time, especially when a quick response to resilience provision is demanded while taking a vast number of stochastic variables into account. Furthermore, it is hard to develop a generalized control scheme that can be applied to any state condition. Fourth, even though the system models and uncertainties are both known, RCs are moving

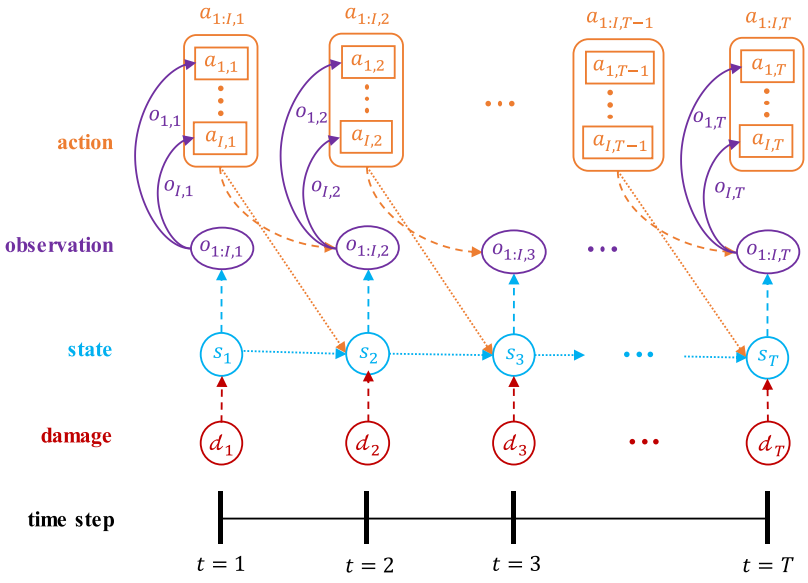


Fig. 3. Architecture of the proposed Dec-POMDP.

towards a decentralized manner that can only observe partial information of the coupled networks. Also, proper incentives presenting the wills and behaviors of RCs may need to be designed to motivate them to participate in the load restoration process [40].

To address the aforementioned challenges, we solve the above problem via a MARL-based approach, which can prevent knowledge exchange among RCs while simultaneously yielding a coordinated scheme for these decentralized agents via appropriate reward function design and rational cooperation mechanism. Additionally, extensive interactions with the environment throughout the learning process can potentially capture system uncertainty. Finally, once the MARL algorithm has been trained thoroughly, the control policies can be deployed in milliseconds for realistic dispatching decisions in response to the resilience problem.

3. Reformulation as a decentralized partially observable markov decision process

3.1. Decentralized partially observable markov decision process

Since each RC is operated in a decentralized manner and can only observe partial information of the power-gas-transportation network, it is reasonable to model the problem of RCs routing and repairing for resilience enhancement as a *Decentralized Partially Observable Markov Decision Process* (Dec-POMDP) [18], of which its architecture is shown in Fig. 3. Specifically, the Dec-POMDP is defined by $\langle I, \mathcal{S}, \mathcal{O}, \mathcal{A}, \mathcal{R}, \mathcal{T}, \gamma \rangle$, which includes I RC agents, a set of global states $s \in \mathcal{S}$, a collection of local observations $\{o_i \in \mathcal{O}_i\}$, a collection of action sets $\{a_i \in \mathcal{A}_i\}$, a collection of reward functions $\{r_i \in \mathcal{R}_i\}$, and a state transition function $\mathcal{T}(s, a_{1:I}, \omega)$ conditioned on the environment state s, all agents' actions $a_{1:I}$, as well as the environment stochasticity ω representing the power-gas-transportation network's uncertain parameters, e.g., traffic volume, electric demand, gas demand, and PV power generation. The time interval between two consecutive time steps $\Delta t = 1$ h.

At time step t, there is a particular state s_t where the environment is in. This state emits a joint observation according to the observation model (dashed blue arrows) from which each agent observes its individual component (indicated by solid purple arrows). Then, based on its local observation $o_{i,t}$, each agent i chooses an action $a_{i,t}$ in accordance with the policy $\pi(a_{i,t}|o_{i,t})$. The combined actions of all agents cause the environment to transit into a new state s_{t+1} according to the transition function $\mathcal{T}$ (dotted orange arrows). Then, each agent i receives a reward $r_{i,t}$ and a new local observation $o_{i,t+1}$. This process continues until each agent i emits a trajectory of observations, actions, and rewards: $\tau_i = o_{i,1}, a_{i,1}, r_{i,1}, o_{i,2}, \ldots, r_{i,T}$ over $\mathcal{O}_i \times \mathcal{A}_i \times \mathcal{O}_i \to \mathbb{R}$. Each agent i seeks to maximize its cumulative discounted reward $R_i = \sum_{t=0}^{T} \gamma^t r_{i,t}$, where $\gamma \in [0, 1)$ is the discount factor and T is 24 time steps in the daily horizon. It is noted that apart from the local information (e.g., located traffic volume, nodal demand and PV generation, and individual repair status), each RC agent i can also observe the system outage status $\mathscr{A}$ (e.g., damaged component and location), which is transmitted by the central monitor. However, the RC agents are still operated in a decentralized manner since the central monitor can only acquire partial information while the agents' local information is private. The detailed components of Dec-POMDP in the proposed problem are detailed in the following subsections:

3.2. Environment state and local observation

The environment state, $s_t = \{o_{1,t}, \ldots, o_{I,t}\} \in \mathcal{S}$, describes the configurations of all RC agents at time step t, where the local observation $o_{i,t}$ of each RC agent i at time step t can be defined as an 11-dimensional vector:

$$o_{i,t} = [R^{rd}_{i,t}, V^{rd}_{i,t}, N^{rd}_{i,t}, S^{dm}_{t}, N^{dm}_{t}, P^{ed}_{i,t}, P^{pv}_{i,t}, G^{gd}_{i,t}, RS^{rc}_{i,t}, RT^{rc}_{i,t}, S^{rc}_{i,t}] \in \mathcal{O}_i, \quad \forall i \in I, \ \forall t \in T, \tag{28}$$

consisting of four parts: (1) the transportation information of the road index $R^{rd}_{i,t}$ and its corresponding traffic volume $V^{rd}_{i,t}$, as well as the terminated node index $N^{rd}_{i,t}$ to which the RC agent i is moving; (2) the outage information of the damaged components' status S^{dm}_{t} and location N^{dm}_{t} in the MEMG, which are broadcast by the central monitor; (3) the power and gas information of the nearby ED $P^{ed}_{i,t}$, GD $G^{gd}_{i,t}$, and PV generation $P^{pv}_{i,t}$; and (4) the repairing information of the time horizon $RT^{rc}_{i,t}$ and resources $RS^{rc}_{i,t}$ required by the RC agent i to repair the damaged component, as well as the RC agent i's current resource status $S^{rc}_{i,t}$.

3.3. Action

The action $a_{i,t}$ executed by each RC agent i at each time step t consists of transportation network routing decisions and power-gas network repair decisions, which are defined as a 2-dimensional vector:

$$a_{i,t} = [a^{trl}_{i,t}, a^{rep}_{i,t}] \in \mathcal{A}_i, \ \forall i \in I, \ \forall t \in T \tag{29}$$

where the discrete action $a_{i,t}^{trl} \in \{0,1,2,3\}$ is selected from the set of four potential routing directions upon the currently located transportation node, where 0 indicates the idle status without any routing behavior and 1, 2, 3 indicates left, right, go forward, respectively. The binary action $a_{i,t}^{rep} \in \{0,1\}$, on the other hand, represents the repairing decision of RC agent i (1 if repairing the component, 0 if not). It is worth noting that these two actions $a_{i,t}^{trl}$ and $a_{i,t}^{rep}$ are mutually exclusive, which means that RC agent i cannot travel in the transportation network while also repairing damaged components in the power-gas network.

3.4. State transition

The state transition from time step t to $t+1$ is governed by function $s_{t+1} = \mathcal{T}(o_{1:I,t}, a_{1:I,t}, \omega_t)$. It is noted that the transition is influenced partly by the agents' actions $a_{1:I,t}$ and partly by the environment's stochasticity ω_t. In the examined problem, this corresponds to the exogenous state features $\omega_t = [V_{i,t}^{rd}, P_{i,t}^{el}, P_{i,t}^{pv}, G_{i,t}^{gl}]$ which are decoupled from the agents' actions and are characterized by inherent variability and uncertainty. In this context, it presents significant challenges to identify suitable probabilistic models which can fully capture such randomness since it is influenced by many exogenous factors, such as driving behaviors, energy usage behaviors, solar radiation, and load distinction. RL, however, remedies this problem in a data-driven approach that does not rely on accurate models of the underlying uncertainties but learns their probability characteristics through the historic data or experience acquired from the environment via machine learning techniques [18].

By contrast, the state transition for endogenous state features $R_{i,t}^{rd}$, $N_{i,t}^{rd}$, S_t^{dm}, N_t^{dm}, and $S_{i,t}^{rc}$ are determined by actions $a_{i,t}^{trl}$ and $a_{i,t}^{rep}$ adopted at time step t. When an RC agent i travels in the transportation network, the vehicle's global position system (GPS) can automatically detect the local map information $R_{i,t}^{rd}$ and $N_{i,t}^{rd}$. Furthermore, when an RC agent i moves to the location of a damaged component N_t^{dm}, the associated repairing time $RT_{i,t}^{rc}$ and resources $RS_{i,t}^{rc}$ can be observed. If the RC agent i decides to repair this component (i.e., $a_{i,t}^{rep} = 1$), the remaining resources $S_{i,t}^{rc}$ at time step t can be updated after the component is repaired:

$$S_{i,t+1}^{rc} = S_{i,t}^{rc} - a_{i,t}^{re} RS_{i,t}^{rc}, \ \forall i \in I, \ \forall t \in T \tag{30}$$

At the same time, the status S_t^{dm} of this damaged component will be transited to 1 from 0; the location N_t^{dm} of damaged component will be updated as well.

3.5. Reward function

At the end of time step t, each RC agent i obtains its reward $r_{i,t}$. The objective of each RC agent i is to minimize the traveling time in the routing process and the load shedding cost. Thus, the reward function can be designed as two parts: (1) the penalty for traveling time in the transportation network; and (2) the reward for providing load restoration through repairing behaviors.

$$r_{i,t} = -\kappa_1 (u_{i,r,t}^{rd} T_{r,t}^{trl}) + \kappa_2 \xi_{i,t}, \ \forall i \in I, \ \forall t \in T, \ \forall r \in R \tag{31}$$

where the binary $u_{i,r,t}^{rd}$ indicates whether (1) or not (0) the RC agent i is on traffic road r at time step t. Furthermore, the index $\xi_{i,t}$ indicates the contribution of each RC agent i to the system overall demand requirement, which can be expressed as:

$$\xi_{i,t} = \frac{|P_{i,t}^{rc}|}{\sum_{d \in ED} P_{d,t}^{ed}} + \frac{|G_{i,t}^{rc}|}{\sum_{d \in GD} G_{d,t}^{gd}}, \ \forall i \in I, \ \forall t \in T \tag{32}$$

where $P_{i,t}^{rc}$ represents the power flow through the repaired line or the power generation supplied by the repaired DG; and $G_{i,t}^{rc}$ represents the gas flow through the repaired pipeline. As a result, it can be observed that the higher value of $\xi_{i,t}$ indicates the better performance of the system restoration condition provided by RC agent i. Finally, κ_1 and κ_2 are two weighting factors that decide on the relative importance of quality terms with respect to each other.

4. Hierarchical multi-agent reinforcement learning

To solve the Dec-POMDP defined above, we propose a novel MARL algorithm named HMAPPO, with its general architecture shown in Fig. 4. HMAPPO derives four concrete implementation details that are insightful and particularly critical to our proposed resilience problem: (1) constructing a hierarchical architecture with a two-level framework [33] to select either a transportation network route or a power-gas network repair; (2) taking advantage of the MAPPO algorithm [34] in sampling efficiency, learning stability, and hyperparameter robustness to optimize the policy; (3) approximating an abstracted state-value function through a set of collective indexes that represent system dynamics to stabilize the multi-agent training performance with privacy protection; (4) using the parameter-sharing (PS) technique [35] to update a single common model for all agents to speed up training process.

To clarify the process of the proposed HMAPPO algorithm in Fig. 4 in detail, we describe six steps inside the algorithm, which can be summarized as below.

Step 1: At time step t, each RC agent i can directly observe its local information of (i) transport road index, traffic volume, and the terminated node index moving to; (ii) nearby PV power generation, electric and gas demand; (iii) current resources as well as the repairing time and resources of the damaged component. In addition, each RC agent i can also acquire the system outage information of the damaged component's status and location transmitted by the central monitor. As a result, the concatenation of local information and system outage information is regarded as the final state o_i observed by each agent i, which is also defined in Eq. (28).

Step 2: In observing the local observation o_i, the RC agent i performs the HL action x_i based on the HL policy $\mu_\psi(x|o)$. The HL action is a binary decision switching between the transportation network and the power-gas network. It is assumed in Fig. 4 that the RC agent i switches to making routing decisions in the transportation network (green solid line) while the decision-making process in the power-gas network (yellow dashed line) is isolated.

Step 3: After selecting the HL action x_i, the RC agent i can perform the corresponding LL action a_i. There are two kinds of LL action: one is for routing action a_i^{trl} in the transportation network if the HL action is switched to the transportation system, and the other one is for repairing action a_i^{rep} in the power-gas network if the HL action is switched to the energy system. Because HMAPPO employs the PS technique, the LL routing policy $\pi_{\phi^{trl}}(a|o)$ and the repairing policy $\pi_{\phi^{rep}}(a|o)$ are shared among all RC agents.

Step 4: Each RC agent i executes its routing action or repairing action to the environment based on the LL policies. It is assumed in Fig. 4 that the RC agent i executes the routing action a_i^{trl} to the transportation network (green solid line). It is noted that the transportation routing action and the energy repairing action cannot happen simultaneously for one RC agent, but can happen simultaneously among different RC agents (i.e., the scenario exists with some RC agents making routing decisions in the transportation network and others making repairing decisions in the power-gas network). The MGCC solves the linear network algorithm for the coupled power-gas network and computes the restoration index ξ_i for each RC agent i. The environment can also provide each RC agent i with new local observation o_i and reward r_i.

Step 5: The central monitor retrieves the restoration index ξ_i from the environment and passes it on to each agent i. Each agent i will then store one experience to its trajectory τ_i += $[o_i, x_i, a_i, r_i, \xi_i]$ that is used to update the network weights.

Step 6: Once a batch of trajectories is fully filled, they can be used to update the critic network as well as the one HL and two LL actor networks.

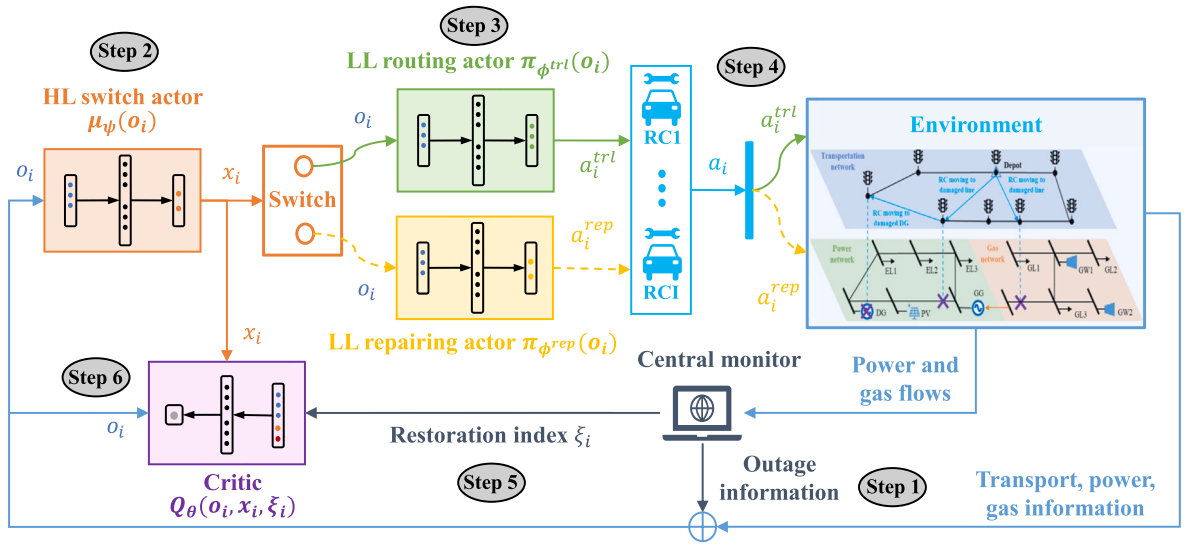


Fig. 4. Architecture of the proposed HMAPPO algorithm.

4.1. Learn two-level hierarchies

Hierarchical reinforcement learning (HRL) is a variant of RL which extends the decision problem to coarser grains with multiple hierarchies [33]. Amongst several HRL algorithms, the two-level framework is a temporal abstraction for RL, where the high-level (HL) action lasts for a number of steps through the low-level (LL) actions. Specifically, for any RC agent i observing $o_{i,t}$ at time step t, an UL action is selected according to the UL policy $x_{i,t} = \mu(x|o) \rightarrow [0,1] \in \mathcal{X}_i$ (e.g., when RC i arrives at the location of a damaged component at time step t, there is a $x_{i,t} = 90\%$ probability to select repairing in power-gas network and the left 10% probability to select routing in transportation network). Afterwards, the LL policy $\pi(a|o)$ is used to determine the LL action $a_{i,t}$ (e.g., repair the damaged component to support resilience). This process is repeated until the UL action switches to routing in transportation network at $x_{i,t} < 50\%$. The reward over the two-level framework is defined as $r_{i,t}$, identical to vanilla RL in (31). Similar to RL on flat actions, the probability transition function over the two-level framework is defined as $p(o_{i,t+1}|o_{i,t}, x_{i,t}) = \sum_{f=1}^{T} p(o_{i,t+1}, f)\gamma^f$, where $p(o_{i,t+1}, f)$ is the probability of a HL action making changes in f time steps. The objective of agent i over two-level framework lasting for f time steps can be rewritten as $R_i(o_{i,t}, x_{i,t}, o_{i,t+f}) = \mathbb{E}[\sum_{z=t}^{t+f} \gamma^{z-t} r_{i,z}]$. Such process continues and then emits a new trajectory of observations, HL action, LL action, and reward for each agent i: $\tau_i = o_{i,1}, x_{i,1}, a_{i,1}, r_{i,1}, o_{i,2}, \ldots, r_{i,T}$ over $\mathcal{O}_i \times \mathcal{X}_i \times \mathcal{A}_i \times \mathcal{O}_i \rightarrow \mathbb{R}$.

In this context, we consider a two-level execution model (as depicted in Fig. 4), in which agent i picks up a HL action $x_{i,t}$ according to its UL policy $\mu(x|o)$ at time step t, then follows the LL policy $\pi(a|o)$ to execute LL actions $a_{i,t}$ for each time step t until the HL action makes changes, at which point this procedure is repeated over time steps T. In order to automatically fit complicated circumstances, we refer to the above hierarchical architecture as an actor–critic architecture [18]. The HL policy $\mu(x|o)$ and LL policy $\pi(a|o)$ belong to the actor part, while a state-value function $V(o, x)$ is introduced as the critic part to specify the expected value of HL selection $x_{i,t}$ in observation $o_{i,t}$. Since the examined problem is characterized by high-dimensions and continuous state space, we assume the actors and critic are all represented by differentiable parameterized function approximators via deep neural networks (DNNs). Specifically, for agent i, the HL policy $\mu_{\psi_i}(x|o)$ is trained by an actor network parameterized by ψ_i; the LL policy $\pi_{\phi_i}(a|o)$ is trained by another actor network parameterized by ϕ_i; and the state value function $V_{\theta_i}(o, x)$ is trained by a critic network parameterized by θ_i.

4.2. Construct categorical policies for routing and repairing actions

Following the selection of HL action $x_{i,t} = \mu_{\psi_i}(x|o)$ for either transportation or power-gas network, the RC agent i needs to make physical actions using the LL policy $a_{i,t} = \pi_{\phi_i}(a|o)$ in observing $o_{i,t}$. Since the routing action $a_{i,t}^{trl}$ and repairing action $a_{i,t}^{rep}$ are both in discrete space, we can adopt the two categorical policies to generate $a_{i,t}^{trl}$ and $a_{i,t}^{rep}$, respectively. In this case, when the HL action $x_{i,t}$ is switched to the transportation network, the RC agent i will perform the routing action $a_{i,t}^{trl}$, and when the HL action $x_{i,t}$ is switched to the power-gas network, it will perform the repairing action $a_{i,t}^{rep}$.

More specifically, we generate two $\texttt{softmax}(f)$ distributions for two discrete actions and compute two actor networks parameterized by ϕ_i^{trl} and ϕ_i^{rep} to output the corresponding probabilities for all potential routing selections and repairing decisions. The optimal actions $a_{i,t}^{trl}$ and $a_{i,t}^{rep}$ are then respectively sampled from these two categorical policies $a_{i,t}^{trl} = \pi_{\phi_i^{trl}}(a|o)$ and $a_{i,t}^{rep} = \pi_{\phi_i^{rep}}(a|o)$ in observation $o_{n,t}$. The two categorical policies, $\pi_{\phi_i^{trl}}$ and $\pi_{\phi_i^{rep}}$, are then separately updated by the MAPPO algorithm [34], which minimizes their respective clipped surrogate objectives to constrain the policy update, as follows:

$$L_{i,t}^{\text{CLIP}}(\phi_i^{trl}) = \hat{\mathbb{E}}_t\left[\min(\zeta_{i,t}^{trl}\hat{A}_{i,t}, \text{clip}(\zeta_{i,t}^{trl}, 1-\epsilon, 1+\epsilon)\hat{A}_{i,t})\right],\ \forall i \in I,\ \forall t \in T \quad (33)$$

$$L_{i,t}^{\text{CLIP}}(\phi_i^{rep}) = \hat{\mathbb{E}}_t\left[\min(\zeta_{i,t}^{rep}\hat{A}_{i,t}, \text{clip}(\zeta_{i,t}^{rep}, 1-\epsilon, 1+\epsilon)\hat{A}_{i,t})\right],\ \forall i \in I,\ \forall t \in T \quad (34)$$

where the first term is the normal policy gradient and the second term trims the policy gradient by clipping the probability ratio $\zeta_{i,t}^{trl}, \zeta_{i,t}^{rep}$ between $[1-\epsilon, 1+\epsilon]$, with ϵ being a small hyperparameter that encourages less per gradient update of the new policy from the old version. The probability ratios $\zeta_{i,t}^{trl}, \zeta_{i,t}^{rep}$ can be expressed as:

$$\zeta_{i,t}^{trl} = \frac{\pi_{\phi_i^{trl}}(a_{i,t}^{trl}|o_{i,t})}{\pi_{\phi_i^{trl}\text{old}}(a_{i,t}^{trl}|o_{i,t})} \quad \text{and} \quad \zeta_{i,t}^{rep} = \frac{\pi_{\phi_i^{rep}}(a_{i,t}^{rep}|o_{i,t})}{\pi_{\phi_i^{rel}\text{old}}(a_{n,t}^{rep}|o_{i,t})},\ \forall i \in I,\ \forall t \in T \quad (35)$$

Then, the HL policy $\mu_{\psi_i}(x|o)$ characterized by discrete actions can be optimized similarly as the above categorical policies in (33)–(34) and the probability ratio of HL policy $\zeta_{i,t}^{x}$ can be derived similarly as the LL discrete probability ratio in (35).

In addition, $\hat{A}_{i,t}$ is a generalized advantage function as:

$$\hat{A}_{i,t} = \delta_{i,t} + \gamma\delta_{i,t+1} + \cdots + \gamma^{T-t+1}\delta_{i,T-1},\ \forall i \in I,\ \forall t \in T \quad (36)$$

Table 1
Network structures and key features for different multi-agent reinforcement learning algorithms.

Algorithm	Actor	Critic	Number of DNNs	Training	Execution	Stationarity	Privacy preserving
CTCE	$\pi(a_{1:I}\|o_{1:I})$	$V(o_{1:I}, x_{1:I})$	4	Centralized	Centralized	Yes	No
DTDE	$\pi(a_i\|o_i)$	$V(o_i, x_i)$	$\|I\|\times 4$	Decentralized	Decentralized	No	Yes
CTDE	$\pi(a_i\|o_i)$	$V(o_{1:I}, x_{1:I})$	$\|I\|\times 4$	Centralized	Decentralized	Yes	No
HMAPPO	$\pi(a_i\|o_i)$	$V(o_i, x_i, \xi_i)$	4	Decentralized	Decentralized	Yes	Yes

$$\delta_{i,t} = r_{i,t} + \gamma V_{\theta_i}(o_{1:I,t+1}, x_{1:I,t+1}) - V_{\theta_i}(o_{1:I,t}, x_{1:I,t}),\ \forall i \in I,\ \forall t \in T \quad (37)$$

where $V_{\theta_i}(o, x)$ is the state-value function taking all agents' local observations $o_{1:I}$ and UL actions $x_{1:I}$ in centralized training [34], which is estimated by a critic network parameterized by θ_i defined in Section 4.1. It is also mentioned that providing all local information to the critic can stabilize learning and foster coordinated behaviors for all agents [34].

4.3. Abstract restoration index

However, the centralized critic network taking all agents' local information may raise problems. First, the shared information among all agents can destroy their privacy, since these private-owned RCs are not willing to exchange their dispatch behaviors with each other [40]. Second, concatenating all local information in centralized training may not contain sufficient global information to reduce a POMDP to an MDP as there can be critical information (e.g., load shedding quantity) which is not observed by any of the agents during training. Third, even if this information can be successfully obtained, it is impractical to apply centralized critic to a large-scale system since the joint information increases proportionally with the agent size. This paper thus assumes the central monitor as a trusted third party who can spread to each agent the extra information representing the system's global dynamics (e.g., load restoration status). As such, we approximate the multi-agent joint state-value function of each agent i as:

$$V_{\theta_i}(o_{1:I}, x_{1:I}) = V_{\theta_i}(o_i, x_i, \xi_i),\ \forall i \in I \quad (38)$$

where ξ_i denotes the contribution of agent i to the system's overall load restoration status in (32). It can be observed that ξ_i is an embedded function that not only abstracts all other agents' local observations (e.g., $P^{el}_{i-,t}, P^{pv}_{i-,t}, G^{gl}_{i-,t}, S^{ln}_{i-,t}, \forall i- \in \mathcal{I}(i-)$, where $\mathcal{I}(i-)$ donates the set of all other agents $i-$ apart from i), but also reflects the status of agent i providing system resilience (the higher value of ξ_i indicates the better performance of contributing to resilience enhancement, and vice versa). As a result of incorporating ξ_i into the state-value function estimation, each agent can make acquainted decisions based on the impact of self and all other agents' observations and actions, despite not knowing their local information and control behaviors, thereby protecting the privacy and also improving the scalability.

It should be noted that $\xi_{i,t}$ for agent i at time step t cannot be directly embedded into the current observation $o_{i,t}$ because it is only available after MGCC solves the linearized network model after receiving all the agents' current actions (which are conditioned on the current observations). However, because the training process is carried out after gaining experience from the environment, the index $\xi_{i,t}$ can be applied to the training process of the critic network.

4.4. Training networks in PS technique

Before implementing the training process, we first discuss the conventional MARL frameworks in the literature and then highlight the motivation of using PS technique in our proposed HMAPPO algorithm. In general, prior works have identified three major paradigms for MARL, including (i) centralized training with centralized execution (CTCE); (ii) decentralized training with decentralized execution (DTDE); and (iii) centralized training with decentralized execution (CTDE), of which their main features are outlined in Table 1. In addition, all the corresponding features of our proposed HMAPPO algorithm are also reported in Table 1.

It can be observed from Table 1 that CTCE takes all agents' local observations as input and computes all agents' actions simultaneously. So, both actor and critic networks are updated in a centralized manner rather than by individual agents. However, the implementation of CTCE may raise agents' opposition, since they are generally unwilling to reveal their private information and exchange such information with each other. DTDE, on the other hand, allows each agent to train its individual actor and critic networks in a decentralized manner that only requires local information, e.g., local observations and actions. As a result, private information can be protected. However, because the RC agents in the power and transport environment can influence each other, a decentralized training framework that recognizes the RC agents as independent entities while ignoring the correlation between them may raise the non-stationary issue in the environment, thereby often leading to an ineffective training performance. To this end, CTDE effectively circumvents the challenge of environmental non-stationarity during the training process by knowing the information of all agents' local observations. During test time, the critic network is not needed, and the policy execution is fully decentralized through each agent's actor network, which only takes its own local observation as input. Nevertheless, as with centralized training, CTDE is not privacy-preserving.

In this context, we would like to propose a novel HMAPPO algorithm, in which the critic network takes the restoration index ξ_i as input apart from the local observation o_i and the HL action o_i. As discussed in Section 4.3, the deployment of the index ξ_i can effectively abstract the system overall status, thereby stabilizing the training performance. Furthermore, the other agents' information is not required to ensure the privacy of the training process. More specifically, during each training iteration, HMAPPO runs for all agents using the shared one HL policy $\mu_\psi(x|o)$ and two LL policies $\pi_{\phi^{trl}}(a|o)$ and $\pi_{\phi^{rep}}(a|o)$ for T time steps and updates them with the collected trajectories τ_i (also inserting index $\xi_{i,1:T}$ transmitted by the central monitor for each trajectory). Once a batch of trajectories are collected from the buffer $\mathcal{J} = \{\tau_\iota\} \sim \mathcal{F}$, the local agents can then make use of them to compute the discounted reward-to-go $\hat{R}_{\iota,t} = \sum_{h=t}^{T} \gamma^{h-t} r_{\iota,h}$ and the advantage function $\hat{A}_{\iota,t}$ based on the shared and abstracted state-value function $V_\theta(o_{\iota,t}, x_{\iota,t}, \xi_{\iota,t})$ for each trajectory ι and time step t. Afterwards, we can train the three actor networks by maximizing their objectives as:

$$\mathcal{L}(\psi) = \frac{1}{J}\sum_{\iota=1}^{J} \min\left(\zeta^{x}_{\iota,t}\hat{A}_{\iota,t}, \mathrm{clip}(\zeta^{x}_{\iota,t}, 1-\epsilon, 1+\epsilon)\hat{A}_{\iota,t}\right), \quad (39)$$

$$\mathcal{L}(\phi^{trl}) = \frac{1}{J}\sum_{\iota=1}^{J} \min\left(\zeta^{trl}_{\iota,t}\hat{A}_{\iota,t}, \mathrm{clip}(\zeta^{trl}_{\iota,t}, 1-\epsilon, 1+\epsilon)\hat{A}_{\iota,t}\right), \quad (40)$$

$$\mathcal{L}(\phi^{rep}) = \frac{1}{J}\sum_{\iota=1}^{J} \min\left(\zeta^{rep}_{\iota,t}\hat{A}_{\iota,t}, \mathrm{clip}(\zeta^{rep}_{\iota,t}, 1-\epsilon, 1+\epsilon)\hat{A}_{\iota,t}\right), \quad (41)$$

here J refers to the batch size. The critic network is trained by minimizing the loss function of mean-squared error

$$\mathcal{L}(\theta) = \frac{1}{J}\sum_{\iota=1}^{J} \min\left(\hat{R}_{\iota,t} - V_\theta(o_{\iota,t}, x_{\iota,t}, \xi_{\iota,t})\right)^2, \quad (42)$$

The following updates are then applied to the network weights, where $\alpha^{\psi}, \alpha^{\phi^{trl}}, \alpha^{\phi^{rep}}, \alpha^{\theta}$ are the learning rates of the gradient ascent/descent algorithm for actor/critic networks.

$$\psi \leftarrow \psi + \alpha^{\psi} \nabla_{\psi} \mathcal{L}(\psi), \tag{43}$$

$$\phi^{trl} \leftarrow \phi^{trl} + \alpha^{\phi^{trl}} \nabla_{\phi^{trl}} \mathcal{L}(\phi^{trl}), \tag{44}$$

$$\phi^{rep} \leftarrow \phi^{rep} + \alpha^{\phi^{rep}} \nabla_{\phi^{rep}} \mathcal{L}(\phi^{rep}), \tag{45}$$

$$\theta \leftarrow \theta + \alpha^{\theta} \nabla_{\theta} \mathcal{L}(\theta), \tag{46}$$

Finally, the pseudo-code of HMAPPO is shown as below:

Algorithm 1 HMAPPO for I agents

Initialize the shared weights $\psi, \phi^{trl}, \phi^{rep}, \theta$ for actor and critic networks
Set learning rates $\alpha^{\psi}, \alpha^{\phi^{trl}}, \alpha^{\phi^{rep}}, \alpha^{\theta}$
for episode (i.e., day) $epi = 1$ to E **do**
  Initialize the global state s_0 and local observation $o_{i,0}$
  For each agent i, sets an empty buffer $\mathcal{F} = \{\}$
  For each agent i, sets an empty trajectory $\tau_i = []$
  For each agent i, selects HL switching action $x_{i,0} = \mu_{\psi}(x|o)$ in observing $o_{i,0}$
  for time step (i.e., 1 h) $t = 1$ to T **do**
    repeat
      for agent (i.e., RC) $i = 1$ to I **do**
        Selects LL routing action $a_{i,t} = a_{i,t}^{trl} = \pi_{\phi^{trl}}(a|o)$ (if HL action is for transportation network) or repairing action $a_{i,t} = a_{i,t}^{rep} = \pi_{\phi^{rep}}(a|o)$ (if HL action is for power-gas network)
      end for
      Execute all agents' actions $a_{1:I,t}$ to the environment
      MGCC solves the linearized network model and collects the load shedding quantity $P_{d,t}^{ls}$ and $G_{d,t}^{ls}$
      for agent (i.e., RC) $i = 1$ to I **do**
        Observes reward $r_{i,t}$ and next observation $o_{i,t+1}$
        Central monitor transmits index $\xi_{i,t}$ to local agent i
        Stores one sample experience to trajectory τ_i += $[o_{i,t}, x_{i,t}, a_{i,t}, r_{i,t}, \xi_{i,t}]$
        while time step t % $J = 0$ **do**
          Agent i collects a set of trajectories τ_i from buffer $\mathcal{F}$, then computes advantage function $\hat{A}_{i,t}$ and discounted reward-to-go $\hat{R}_{i,t}$
          Updates network weights $\psi, \phi^{trl}, \phi^{rep}, \theta$ in (43)-(46)
        end while
      end for
      Update local observation $o_{i,t} \leftarrow o_{i,t+1}$
    until UL action $x_{i,t} = \mu_{\psi}(x|o)$ is switched in new observation $o_{i,t}$
    Update HL action $x_{i,t} \leftarrow x_{i,t+1}$
  end for
end for

4.5. Test process

As stated in Section 4.4, the HMAPPO training process takes E episodes until the trained policy is converged. In the test process, we firstly collect the weight parameters ψ of the HL actor network and ϕ^{trl}, ϕ^{rep} of the two LL actor networks trained in Algorithm 1. For each time step t in the test day d, each RC agent i observes the local observation $o_{i,t}$, then firstly selects the HL action $x_{i,t} = \mu_{\psi}(x|o)$ and secondly executes the LL action $a_{i,t}^{trl} = \pi_{\phi_i^{trl}}(a|o)$ (if $x_{i,t} < 50\%$ when HL action is switched to the transport network) or $a_{i,t}^{rep} = \pi_{\phi_i^{rep}}(a|o)$ (if $x_{i,t} \geq 50\%$ when HL action is switched to the power-gas network) to the environment. The LL actions of all RC agents are then mapped to the operation model of power-gas-transport network (environment), transiting to the next state s_{t+1} and local observation $o_{i,t+1}$ (Section 3.4). Each RC agent i can also obtain its own reward $r_{i,t}$ (Section 3.5). Finally, the pseudo-code of the test process is shown as below:

Algorithm 2 Test process of HMAPPO

Load the weight parameters ψ, ϕ^{trl}, ϕ^{rep} trained by Algorithm 1
for test day = 1 : D **do**
  Initialize the global state s_0 and local observation $o_{i,0}$
  Selects HL switching action $x_{i,0} = \mu_{\psi}(x|o)$ in observing $o_{i,0}$
  for time step (i.e., 1 h) $t = 1$ to T **do**
    repeat
      for agent (i.e., RC) $i = 1$ to I **do**
        Selects LL routing action $a_{n,t} = a_{i,t}^{trl} = \pi_{\phi^{trl}}(a|o)$ (if HL action is for transportation network) or repairing action $a_{n,t} = a_{i,t}^{rep} = \pi_{\phi^{rep}}(a|o)$ (if HL action is for power-gas network)
      end for
      Execute all agents' actions $a_{1:I,t}$ to the environment
      MGCC solves the linearized network model and collects the load shedding quantity $P_{d,t}^{ls}, G_{d,t}^{ls}$
      for agent (i.e., RC) $i = 1$ to I **do**
        Observes reward $r_{i,t}$ and next observation $o_{i,t+1}$
      end for
      Update local observation $o_{i,t} \leftarrow o_{i,t+1}$
    until HL action $x_{i,t} = \mu_{\psi}(x|o)$ is switched in new observation $o_{i,t}$
  end for
end for

5. Input data and experiment setup

5.1. Experiment setup

The examined MEMG includes a 6-bus meshed power network and a 6-node radial gas network modified from [41], as illustrated in Fig. 5. DERs, including 1 DG, 1 GG, 1 PV, and 2 GWs, are appropriately deployed in the power and gas networks, respectively. Load distinction into 2 essential loads (ED1 and GD1) with high shedding costs and 4 non-essential loads (ED2, ED3, GD2, and GD3) with low shedding costs is located in the electric and gas buses. The load shedding costs for essential and non-essential EDs are 2.5 £/kWh and 1.5 £/kWh, respectively [30]. The load shedding costs for essential and non-essential GDs are 2.0 £/S m^3 h and 1.5 £/S m^3 h, respectively. The operation data of ED and PV generation in the power system is obtained from a real-world open-source dataset Ausgrid [42]. The operation data of GD in the gas network is obtained from [41]. In order to capture uncertainties associated with PV generation and demand profiles, we select the PV and load data from July 2012 to May 2013 (17 months) as the training set and the PV and load data in June 2013 (1 month) as the test set. Finally, the technical parameters of 1 DG, 1 GG, 2 GWs, and 3 RCs are presented in Table 2. Here, we consider three kinds of RCs with different repair abilities.

The transportation network is a 9-node and 12-edge road map as shown in Fig. 6. It is assumed that there are 8 potential damaged components in the studied MEMG as depicted in Fig. 5, which includes 1 DG, 1 PV, 3 distribution lines in the power network, and 3 pipeline lines in the gas network. Their corresponding resources and periods for restoration are presented in Table 4. The traffic data of the transportation network is presented in Table 3, expressing the transportation time without congestion and the distance between two adjacent nodes. It can be further observed from Table 3 that each road is characterized by its distance and travel time to simulate the dynamic congestion impact during the RC routing process. In this regard, it can be anticipated that the RC agents will choose a more strategic routing scheme with shorter transport time to save more time for connecting with the network and obtaining higher rewards by providing resilience. Similar to the uncertainties of PV generation and demand in the energy system, the uncertainties associated with the traffic volumes in the transportation network are also characterized by the real-world dataset, which is collected from [43].

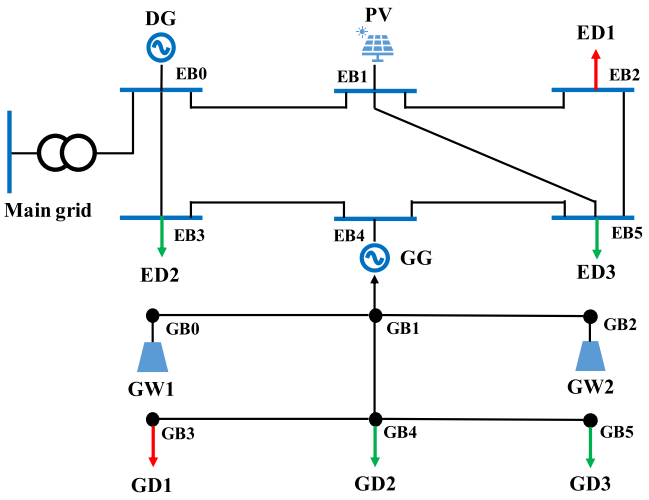


Fig. 5. MEMG of 6-bus power and 6-node gas network.

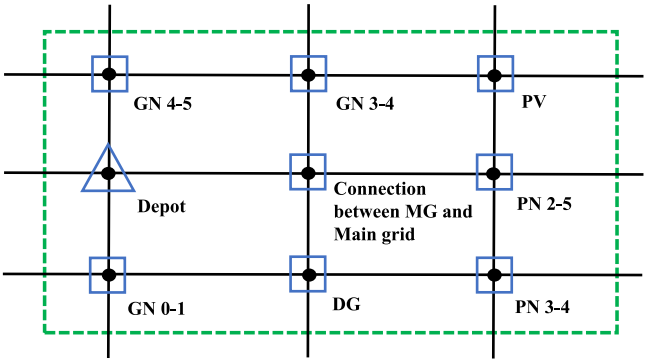


Fig. 6. Transportation network of 9-node and 12-edge map.

Table 2
Technical parameters of DG, GG, GW and RC.

Component	Parameters	Values
DG	$\underline{P}^{dg}$, $\overline{P}^{dg}$, $\underline{Q}^{dg}$, $\overline{Q}^{dg}$	0 kW, 150 kW, −67 kVAR, 150 kVAR
GG	$\underline{P}^{gg}$, $\overline{P}^{gg}$, $\underline{Q}^{gg}$, $\overline{Q}^{gg}$, b^{gg}	0 kW, 200 kW, −100 kVAR, 200 kVAR, 0.5 Sm^3/kW
GW1	$\underline{W}^{gw}$, $\overline{W}^{gw}$	0 Sm^3, 300 Sm^3
GW2	$\underline{W}^{gw}$, $\overline{W}^{gw}$	0 Sm^3, 200 Sm^3
RC1	$\overline{S}^{rc}$	6 units
RC2	$\overline{S}^{rc}$	8 units
RC3	$\overline{S}^{rc}$	10 units

Table 3
Road data of 9-node and 12-edge transportation network.

Road	From node	To node	Travel time (min)	Distance (km)
0	6	7	40.0	46.7
1	7	8	41.7	48.7
2	3	4	52.5	61.3
3	4	5	47.3	55.2
4	0	1	57.0	66.5
5	1	2	39.2	45.7
6	6	3	44.2	51.6
7	7	4	36.3	42.4
8	8	5	47.8	55.8
9	3	0	40.0	46.7
10	4	1	36.7	42.8
11	5	2	38.4	44.8

Table 4
Required resources and time period to restore potential damages.

Potential damage	RS (unit)	RT (h)
Power line (PL) 2 – 5	3	3
Power line (PL) 3 – 4	2	2
Connection line (Grid)	3	4
DG	5	4
PV	2	3
Gas pipeline (GL) 0 – 1	3	3
Gas pipeline (GL) 3 – 4	4	3
Gas pipeline (GL) 4 – 5	3	2

5.2. Benchmarks

In order to validate the superior performance of our proposed HMAPPO in the examined RCs resilience enhancement problem, we compare it against two state-of-the-art MARL algorithms (HPPO and MAPPO), one heuristic algorithm (GA), and one model-based optimization approach (MPC):

(1) HPPO: each RC agent adopts an independent MARL method employing a hierarchical architecture and the PPO algorithm [44]. In this setting, the PS technique and the index ξ are not used. The inputs of critic network are local observation o and HL action x. In other words, the benefit of collective training framework and the key information ξ capturing system dynamics are not considered in this algorithm.

(2) MAPPO: based on our proposed HMAPPO, the RC agents remove the hierarchical architecture while routing and repairing actions are made simultaneously for each time step. In this setting, only one actor network is constructed in MAPPO, and this single actor network inputs the local observation o and outputs the actions of both routing a^{trl} and repairing a^{rep}. The inputs of critic network are local observation o and restoration index ξ. The PS technique is not used.

(3) GA: genetic algorithm (GA) [45] is a kind of heuristic algorithm that defines a set of methods inspired by natural selection such as mutation, inheritance, and crossover. To apply GA to the examined RCs' resilience enhancement problem, we first create a daily genetic representation of the action domain (i.e., routing a^{trl} and repairing a^{rep}) and then define the reward function r as the fitness function. For each time step t, by iterating the fitness function for each generation, GA produces a population of candidate solutions for each RC agent and then selects the optimal solution. The termination criterion is met either when the fitness level does not change in the past 10 populations, or the maximum generations have been reached. It should be noted that GA is also a model-free method since RC agents search for the optimal solutions only based on the action domain and the fitness function, but without knowing the specific power-gas-transportation network.

(4) MPC: model predictive control (MPC) [46] is an advanced optimization method that allows the current time step to be optimized while taking future time steps into account and satisfying a set of constraints. To apply MPC to the examined RCs' resilience enhancement problem, a centralized optimization is constructed with the objective function (7) and constraints (1)–(6), (9)–(27), which assumes the perfect information of the transportation network, power-gas network, RC models, and all technical parameters.

5.3. Implementations of MARL algorithms

The detailed specifications of actor and network networks for three MARL algorithms are presented in Table 5. There are three actor networks in HMAPPO and HPPO, but only one actor network in MAPPO.

Table 5
The general specifications of three examined MARL algorithms.

Mechanism	Actor network	Critic network
HMAPPO	linear(o_dim, 64) → ReLU() → softmax(64, 2) linear(o_dim, 64) → ReLU() → softmax(64, 4) linear(o_dim, 64) → ReLU() → softmax(64, 2)	linear(o_dim+1+1, 64) → ReLU() → linear(64, 1)
HPPO	linear(o_dim, 64) → ReLU() → softmax(64, 2) linear(o_dim, 64) → ReLU() → softmax(64, 4) linear(o_dim, 64) → ReLU() → softmax(64, 2)	linear(o_dim+1, 64) → ReLU() → linear(64, 1)
MAPPO	linear(o_dim, 64) → ReLU() → softmax(64, 8)	linear(o_dim+1, 64) → ReLU() → linear(64, 1)

Table 6
Computational performance (averaged values) of 3 RCs for different MARL algorithms.

Method	Episodic training time (s.)	Number of episodes (#)	Total training time (h.)
HMAPPO	2.98	1433	1.18
MAPPO	1.83	2267	1.15
HPPO	2.82	3000[a]	2.35[a]

[a]Fail to reach convergence within 3000 episodes.

Table 7
Averaged load shedding quantity and cost over 30 test days for different MARL algorithms, GA and MPC.

Method	Power quantity (kWh)	Gas quantity (S m^3 h)	Cost (£)
HMAPPO	568	934	2734
MAPPO	691	1089	3709
HPPO	932	1469	4911
GA	631	1026	3358
MPC	698	1072	3690

The inputs of all actor networks are local observations in O_DIM = 11 dimensions. However, the outputs of actor networks vary for different algorithms. In HMAPPO and HPPO, the outputs of three actor networks are 2 dimensions' HL action x (selecting transportation or energy networks), 4 dimensions' LL routing action a^{trl} (selecting idle, left, right or go forward), and 2 dimensions' LL repairing action a^{rep} (repair or not), respectively. In MAPPO, the outputs of one actor network are 8-dimensional routing and repairing actions $a^{trl} \cup a^{rep}$ (the permutations of 4 routing actions and 2 repairing actions). For the critic network, a linear activation function is used for the output layer, while its input varies for different algorithms. HMAPPO, as the most complex algorithm, inputs the local observation o, HL action x, and restoration index ξ; HPPO and MAPPO, respectively, remove the restoration index ξ and HL action x. Finally, in all three MARL algorithms, we construct one hidden layer with 64 units using RELU as the activation function for all actor and critic networks.

To make the experiments comparable, we run 3000 episodes with $T = 24$ time steps for all MARL algorithms to evaluate their training performance with the same 10 random seeds for the environment and weights initialization. During the training process, we use the Adam optimizer [47] for all actor and critic networks with a learning rate $\alpha^{\psi} = \alpha^{\phi^{trl}} = \alpha^{\phi^{rep}} = 10^{-4}$ and $\alpha^{\theta} = 10^{-3}$, respectively. The batch size $J = 24$ refers to the number of collected trajectories per episode for updating networks. We employ a clip rate $\epsilon = 0.2$ and a discount rate $\gamma = 0.9$ used to expect a long-term return within a trading day of 24 time steps.

6. Case studies

6.1. Training and test performance

This section aims at comparing the training and test performance of three examined MARL algorithms. Fig. 7 illustrates the evolution of episodic reward of 3 examined RCs over 3000 episodes for different MARL algorithms, where the solid lines and the shaded areas respectively depict the moving average over 50 episodes and the oscillations of the reward during the training process. Furthermore, their corresponding averaged episodic training time as well as the averaged number of episodes and averaged total training time required to reach convergence of 3 RCs are collected in Table 6. Finally, the test performance of system load shedding quantity and cost (7) for three MARL algorithms, GA, and MPC over the 30 test days (June 2013) are also compared in Table 7.

The first observation from Fig. 7 is that three MARL algorithms exhibit an increasing trend in reward level for all three RCs at the beginning of the learning process. However, the oscillation of HPPO (green) is very significant. This phenomenon is particularly serious for RC2 and RC3, as both of them fail to reach convergence within 3000 episodes and even exhibit a downward trend. We believe that such an instability issue is mainly driven by independent learning that focuses on local information only while ignoring the system dynamics, thus rendering the environment non-stationary. To this end, MAPPO (orange) involves the restoration index ξ into the critic network that helps learn the system dynamics. It is observed that MAPPO can effectively mitigate such non-stationarity issues and exhibits better performance in stability. However, MAPPO suffers from poor learning efficiency, which results in a slow convergence speed. This is because, lacking the hierarchical architecture, RC agents simultaneously generating routing and repairing actions may lead to inefficient learning performance of the critic network, since one of these two actions becomes meaningless in the environment. Nevertheless, the network still needs to be trained given this uninformative action and its corresponding reward. In this context, our proposed HMAPPO (blue) addresses the aforementioned issues by (1) learning system dynamics through the restoration index ξ and (2) constructing a more reasonable hierarchical architecture to take adaptive options to the environmental status (i.e., performing in either the transportation network or the power-gas network).

We further assess how well they perform computationally during the training procedure. Table 6 shows that MAPPO has the shortest episodic training time (since it only requires training one actor network to compute both routing and repairing actions, eliminating the need for hierarchical architecture), followed by HPPO and HMAPPO (since these two algorithms construct the same actor network and similar critic networks, with the exception that HMAPPO requires the index ξ as an additional input feature for the critic network). Additionally, we see that HMAPPO (around 1433 episodes) demonstrates a faster rate of convergence than MAPPO (around 2267 episodes). This is due to the PS technique updating a single shared model for all agents that have access to a faster algorithm. Due to its instability problem, HPPO fails to reach to obtain convergence within 3000 episodes. Finally, our proposed HMAPPO (1.18 h) costs the similar computational time to MAPPO (1.15 h) but obtains a better policy quality.

Once the learned policies of three MARL algorithms are finished, we collect the learned models and their associated weights from actor networks and then broadcast them to 3 RCs to execute routing and repair actions to the examined power-gas-transportation network over the 30 test days (Algorithm 2). It can be observed from Table 7 that HMAPPO still exhibits the best performance in terms of the lowest (power and gas) load shedding quantity and cost among all three MARL algorithms. Furthermore, we also apply the other benchmarks,

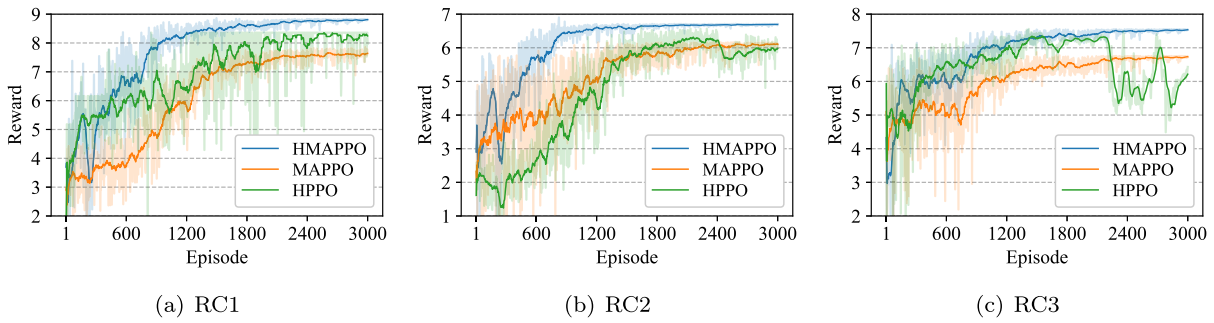


Fig. 7. Episodic reward of 3 RCs (a–c) over 3000 training episodes for different MARL algorithms.

GA and MPC, to solve the problem. It can be observed that even though GA is able to reach a close solution to our proposed HMAPPO, it is computationally expensive and needs to find the solutions for each test day. On the other hand, MPC also exhibits worse performance than our proposed HMAPPO. This is because MPC solves the problem by taking various system uncertainties into account that may converge towards a local optimum. Overall, HMAPPO obtains the lowest load shedding cost among all examined control methods, which is 26.29%, 44.33%, 18.58%, and 25.91% less than MAPPO, HPPO, GA, and MPC, respectively (Table 7).

6.2. Evaluation of RC routing and repairing characteristics

This section aims at analyzing the 3 RCs' routing and repairing characteristics for system resilience enhancement. One outage scenario is appropriately considered for presentation, including two power line outages (PL2 – 5 and PL3 – 4), the connection with the main grid, two gas pipeline faults (GL3 – 4 and GL4 – 5), and 1 power source damage (DG at EB0), as depicted in Fig. 8. The detailed results of 3 RCs' routing and repairing characteristics under this considered outage scenario are illustrated in Figs. 8 and 9. Fig. 10 presents the 3 RCs' source status for repairing the damaged components within the day. Figs. 11 and 12 respectively illustrate the daily power flows of 8 power lines and gas flows of 6 gas pipelines in the coupled power-gas network.

It can be observed from Fig. 8 that all 3 RC agents are operating to repair the damaged components within the day. Specifically, RC1 (blue) leaves the depot and arrives at the damaged main grid connection, then starts the repair at 2:01 and finishes at 6:00. In this case, the main grid returns to normal operation at 6:01, which can also be observed in Fig. 11. Once the main grid connection is fully repaired, RC1 immediately travels to the damaged PL2 – 5 and spends 1 h (6:01–7:00) on the journey. After 3 h' repairing (7:01–10:00), PL2 – 5 returns to normal operation and starts supplying the power system. In contrast to RC1, RC2 (red) chooses to first repair the gas system (i.e., GL3 – 4) at 3:01 and finishes at 6:00. Then, RC2 immediately travels to the damaged PL3 – 4 (arriving at 9:00) and spends 2 h (9:01–11:00) repairing it. As a result, GL3–4 and PL3–4 sequentially return to the normal operation at 6:01 and 11:01, as depicted in Figs. 12 and 11, respectively. Finally, RC3 (green) makes a long joinery (3 h between 4:01–7:00) across the entire power-gas network, spends 2 h (2:01–4:00) and 5 h (7:01–12:00) repairing the damaged GL4 – 5 and DG, respectively. As a result, the corresponding gas flow of GL4 – 5 and the power supply of DG respectively become operable at 4:01 and 12:01, as depicted in Figs. 12 and 11. In the above analysis of the routing and repair processes, it can be highlighted that the gas network and the main grid connection are prioritized to be repaired (in the first round of repair in Fig. 9), which are more important for enhancing overall system resilience. This is because (1) the examined gas network has a radial topology, which is more vulnerable to extreme events compared to the power network having a meshed structure; and (2) repairing the connection with the main grid can provide a large amount of power supply outside the MEMG, further improving the system reliability. Finally, it can be observed from Fig. 10 that all three RCs have almost run out of their resources by the end of the day.

Table 8
Load shedding quantity and cost for idle and HMAPPO dispatch strategies.

Strategy	Idle	HMAPPO
Power Quantity (kWh)	2924	503
Gas Quantity (S m^3 h)	5113	858
Cost (£)	14,190	2490

Go further, we try to evaluate the benefits of RCs' routing and repairing characteristics in providing resilience enhancement by comparing it to the case if RCs are idle (no repairing behavior) under the damaged power-gas system. Specifically, Figs. 13–16 illustrate the power supplies and gas balances of the energy system with and without RCs' dispatches, respectively. It can be observed from Figs. 13–14 that there is only limited load shedding in both power and gas networks after the 3 RCs' first-round repairs (the first 6 h). Then, all the EDs and GDs can be fully supplied once the RCs finish the second-round repair at 10:00. All resources (including main grid, PV, DG, GG, GWs) are utilized for energy supplies. It is noted that the power supply of GG in the power system (in Fig. 13) is supported by the GWs in the gas network (in Fig. 14), considering the energy coefficient for gas consumption in GG. However, it can be found in Figs. 13–14 that there is a large amount of load shedding in both power and gas networks during the scheduling horizon if RCs become idle without any dispatch behaviors. In the power network, the power supplies from the main grid and DG are completely zero. The support from GG is also limited by the damaged power lines. In the gas network, only GW1 is operated to supply the gas system, which is very weak to make the system operate reliably.

Last but not least, the load (3 EDs and 3 GDs) level comparisons between with and without RCs' dispatch in power and gas networks are illustrated in Figs. 17 and 18, respectively. On the one hand, there are different levels of load shedding for all 3 EDs in the power network before RCs finish repairing the connection with the main grid and two damaged PLs. On the other hand, GD1 and GD3 in the gas network have a large amount of load shedding before the GLs are repaired, while GD2 is not influenced by the extreme event since the connection between GD2 and the gas supply is intact. Furthermore, we can also observe that each ED and GD (apart from GD2) exhibits a certain level of load shedding if there are no dispatch behaviors in the restoration process, of which ED2, GD1, and GD3 are particularly serious without any load supply. To make a detailed comparison, we present the load shedding quantity and cost in both power and gas systems for Idle and HMAPPO algorithms in Table 8. It can be observed that HMAPPO can provide 2421 kWh power restoration and 4255 S m^3 h gas restoration over the load supplies in the idle strategy, which results in an 11,700 £ total saving in load shedding cost.

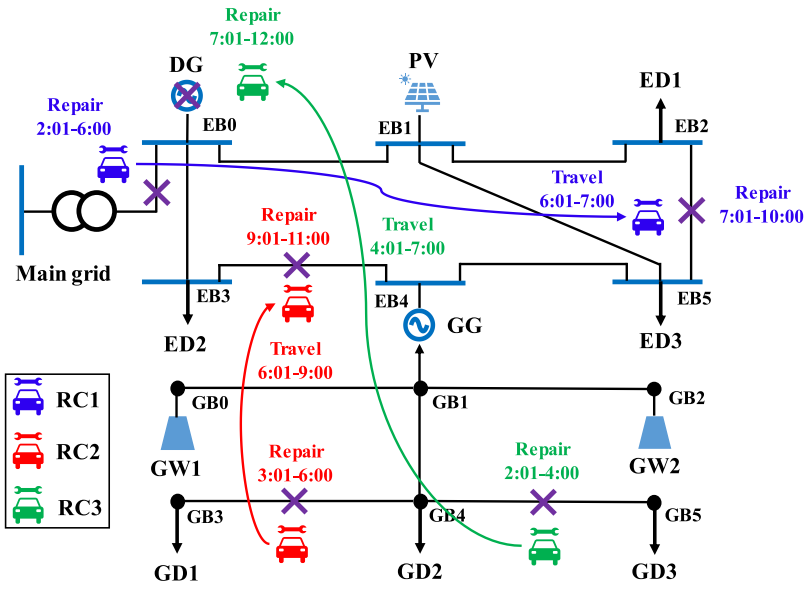


Fig. 8. Routing and repairing characteristics of 3 RCs inside the 6-bus power and 6-bus gas network.

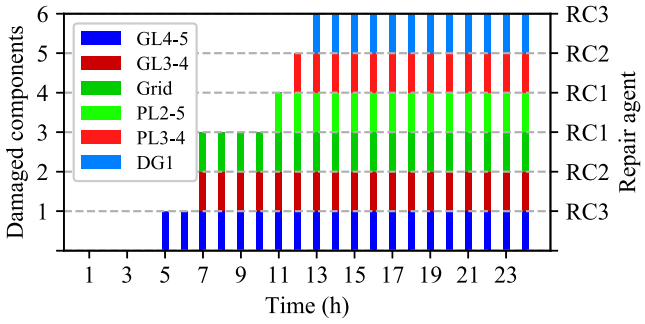


Fig. 9. Damaged components repaired by 3 RCs.

6.3. Scalability analysis in 33-bus power and 15-bus gas system

To further investigate the scalability of the proposed HMAPPO, a larger operation system (a 33-bus power network, a 15-bus gas network, and an 18-node 27-edge transportation network) is utilized in this subsection, which includes 3 DGs, 6 PVs, 2 GGs, 4 GWs, 21 EDs, and 9 GDs, while 8 RCs are employed for providing system resilience. It is worth noting that the studied RC dispatch problem focuses on the distribution network level, which is normally characterized by a radial structure [48]; thus, the modified 33-bus power network is employed in this section for scalability analysis. Furthermore, to capture the severe damage caused by a natural disaster, one outage scenario including 8 PLs, 5 GLs, 1 DG, and 1 GG is appropriately selected for presentation. Fig. 19 illustrates the network structure of the 33-bus power and 15-bus gas network, and also expresses the routing and repairing characteristics of these 8 RCs for resilience enhancement.

First of all, it can be observed from Fig. 19 that all 8 RCs are dispatched to repair the damaged components in a coordinated manner, while most of them perform multiple repairing tasks that fully explore their mobility and flexibility for resilience enhancement. Although the dispatching process of these 8 RCs is very complex, it is not hard to find that network lines are more important for system resilience, which has a higher priority to be repaired. Specifically, RC1 (blue) repairs GL4–5 at 3:01–5:00; RC2 (red) repairs PL7–8 at 2:01–4:00; RC3 (green) repairs PL25 – 26 at 3:01–5:00; RC4 (gray) repairs PL1 – 18 at 2:01–6:00; RC5 (pink) repairs PL0 – 1 (connection with the main grid) at 3:01–6:00; RC6 (brown) repairs PL31 – 32 at 2:01–6:00; RC7 (orange) repairs GL9 – 10 at 2:01-5:00; and RC8 (purple) repairs GL11 – 12 at 3:01–7:00. As a result, all 8 RCs contribute to the network lines' repair in the first-round dispatch, bringing five of the eight PLs and three of the five GLs back to normal operation. Afterward, the other three PLs and two GLs are successfully repaired in the second-round dispatch. Finally, the damaged DG2 and GG2 are repaired by RC2 in the third-round dispatch at 9:01–12:00 and RC7 in the second-round dispatch at 6:01–13:00, respectively.

Figs. 20 and 21 present the load restoration and the corresponding energy supplies in power network and energy balances in gas network, respectively. It can be observed that the load shedding for both power and gas networks occurs at the beginning of the day, since RCs are repairing the damaged components. Once part of these damaged components is fully repaired and return to normal operation around 7:00, all EDs and GDs are completely restored for the rest of the hours within the day. Overall, there are 21.65 MWh of load restoration in the power network and 26.61 M S m^3 h of load restoration in the gas network over the day, which is 92.27% and 95.39% of baseline load, respectively. In this context, we can conclude that the proposed HMAPPO exhibits good scalability in both network and agent sizes, and also shows its effectiveness in system resilience enhancement for the coupled 33-bus power and 15-bus gas network.

7. Conclusions, discussions, and future work

This paper has developed a novel MARL algorithm named HMAPPO for the real-time automatic routing and repairing problem of multiple RCs in a coupled power-gas-transportation network, with the objective of MEMG system resilience enhancement. In detail, a traffic network model is proposed to capture the impact of traffic time and congestion, while a linearized network model is employed to capture the coupling operation of power and gas networks. The proposed HMAPPO has various advantages for this particular problem. First, a hierarchical architecture with a two-level framework is proposed to switch the decision-making process between routing in the transportation network and repairing in the power-gas network. Second, a PS-based MAPPO algorithm is applied to enhance the learning efficiency and also ensure stability and scalability by incorporating the restoration index into the critic network.

Extensive case studies on a real-world dataset have been conducted in two power-gas-transport networks to evaluate the training and test performance of the proposed HMAPPO as well as analyze the dispatch

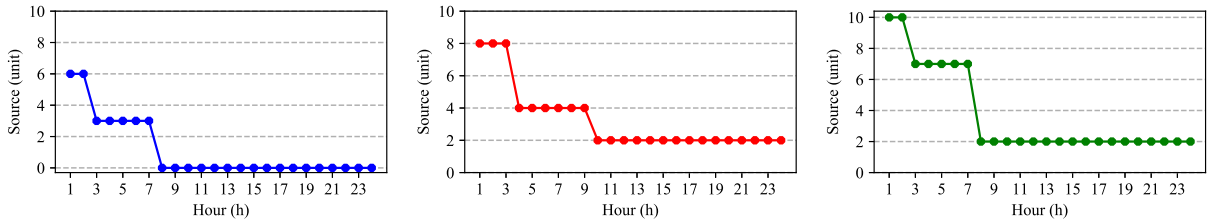


Fig. 10. Source status of 3 RCs for repairing process.

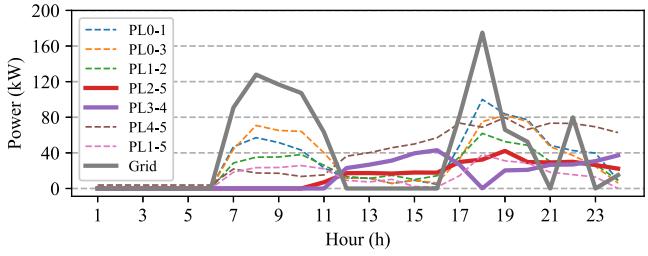


Fig. 11. Power flows in power network.

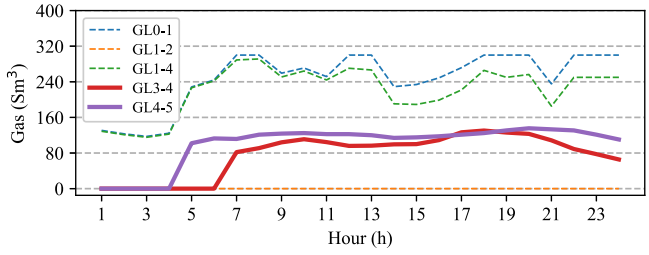


Fig. 12. Gas flows in gas network.

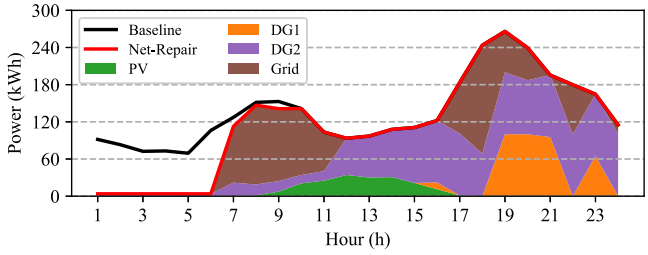


Fig. 13. Energy supplies in power network after repairing.

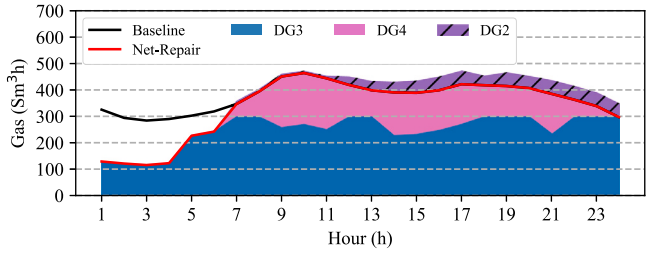


Fig. 14. Energy balances in gas network after repairing.

behaviors of RC agents to provide system resilience. We now discuss the key physical insights observed from the experiment results as below:

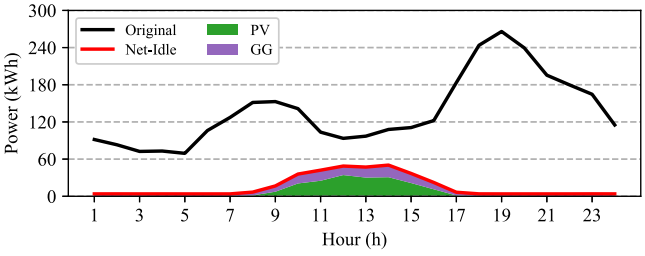


Fig. 15. Energy supplies in power network if no repairing.

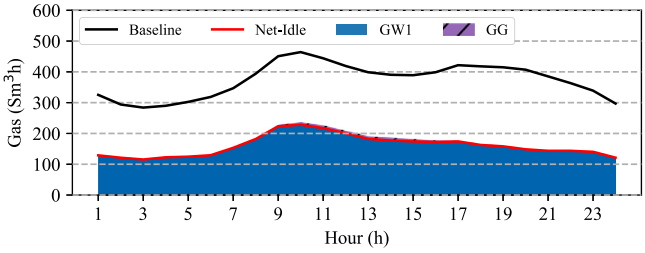


Fig. 16. Energy balances in gas network if no repairing.

(1) The proposed HMAPPO has demonstrated its superior performance in terms of policy performance, stability, and convergence speed compared to the benchmark MARL algorithms during the training process. In addition, HMAPPO in the test process achieves 26.29%, 44.33%, 18.58%, and 25.91% lower averaged system load shedding cost than the benchmark MAPPO, HPPO, GA, and MPC, respectively, over the 30 test days. This is because HMAPPO featuring a hierarchical architecture with a two-level framework is able to take more targeted actions, i.e., select either a transportation network route or a power-gas network repair in observing different state conditions. Furthermore, HMAPPO approximates an abstracted state-value function through a set of collective indexes that can represent system dynamics to stabilize the multi-agent training performance. However, MAPPO and HPPO either do not benefit from the hierarchical architecture or do not benefit from the abstracted state-value function, thereby suffering from the low-quality policy issue and the instability issue, respectively. Finally, HMAPPO learns a faster policy than MAPPO and HPPO due to its parameter-sharing technique that can update a single common policy for all RC agents, speeding up the training process.

(2) The routing and repair behaviors of RCs have been analyzed in the context of a 6-bus power network, a 6-node gas network, and a 9-node 12-edge transportation network to show the effectiveness of learned policy in providing system resilience. In general, the 3 RCs that are utilized can cooperatively repair the damaged lines sequentially within the day and then restore both the power and gas systems to normal operation in 6 and 7 h,

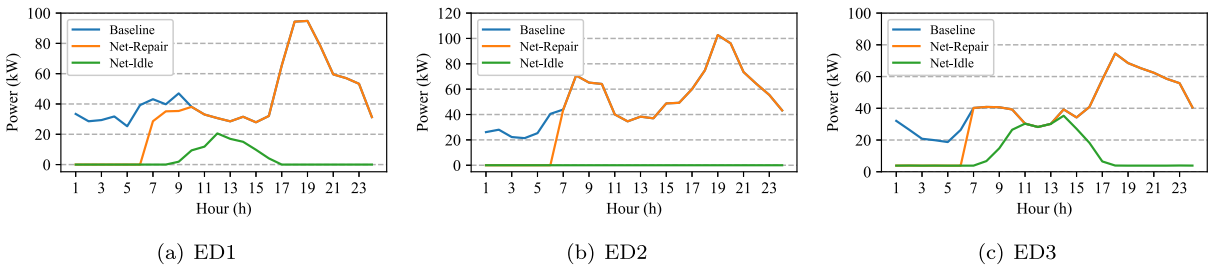


Fig. 17. Load restorations of 3 EDs with and without repairing.

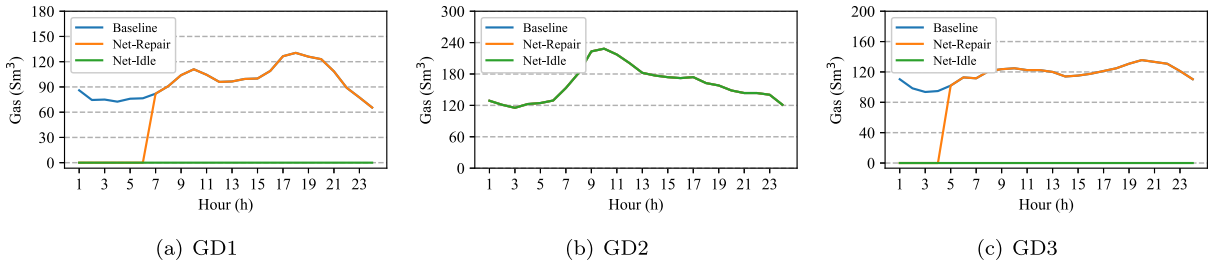


Fig. 18. Load restorations of 3 GDs with and without repairing.

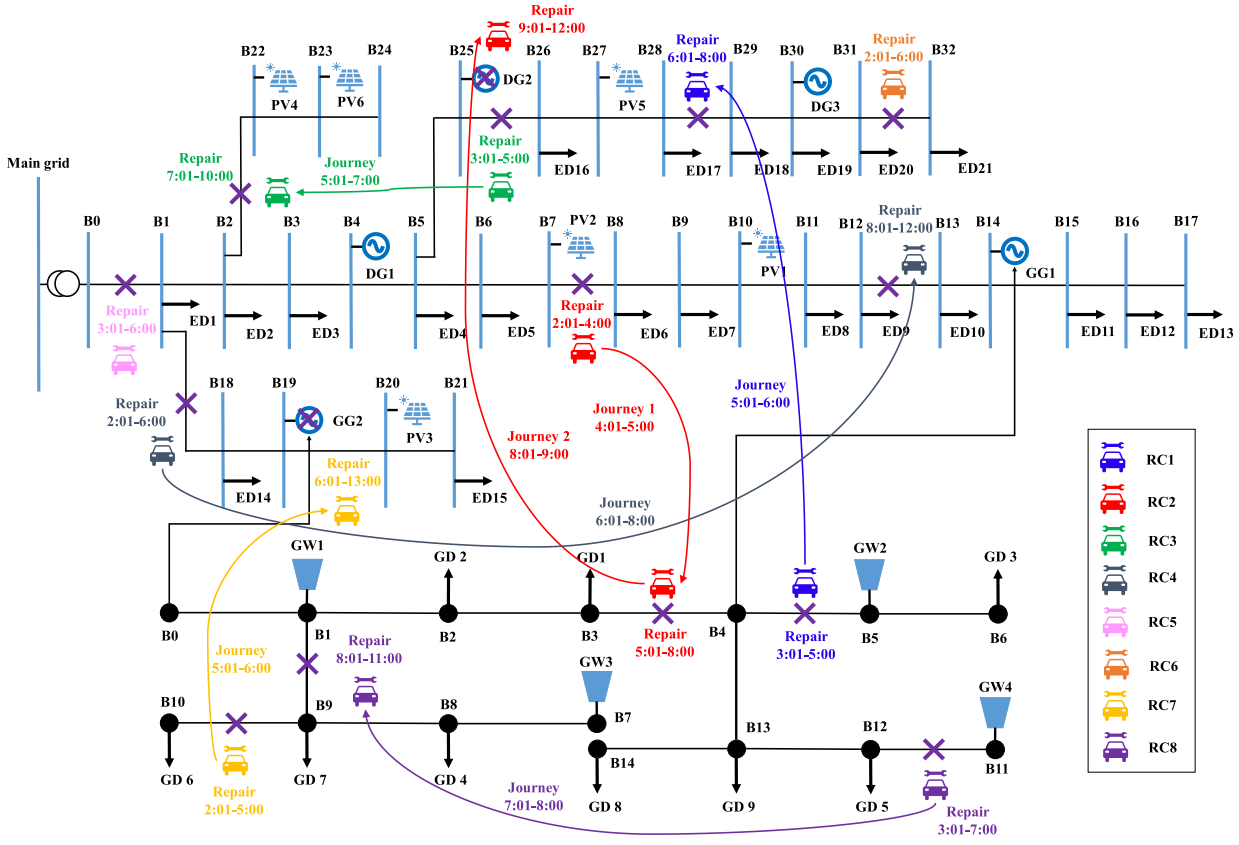


Fig. 19. Routing and repairing characteristics of 8 RCs inside the 33-bus power and 15-node gas network.

respectively. This is because RCs have learned to prioritize the restoration of critical lines. For example, the first task for RC1 is to repair the main grid connection (Fig. 8) in order to allow a more sufficient power supply from the external grid (Fig. 13). Overall, 3 RCs under the proposed HMAPPO algorithm can help the MEMG provide 2421 kWh of power load restoration and 4255 S m^3 h of gas load restoration, which together result in an 11,700 £ total saving in load shedding cost.

(3) The scalability of the proposed HMAPPO algorithm has been evaluated in the context of a 33-bus radial power network, a 15-node radial gas network, and an 18-node 27-edge transportation network, in which 8 RCs are operating to provide system resilience. Owing to the parameter-sharing technique, these 8 RC agents can train a common policy together, which can accelerate the training speed. In addition, the abstracted state-value function inputs the information of local observations,

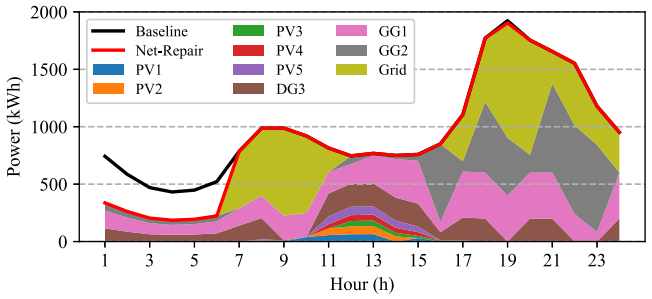


Fig. 20. Energy supplies in power network after repairing.

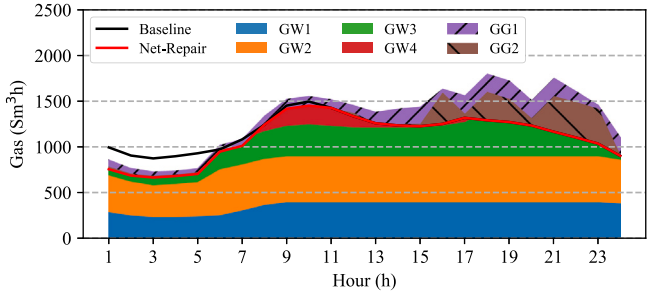


Fig. 21. Energy balances in gas network after repairing.

high-level switch action, and a restoration index that is scalable for multi-agent setup, since the input dimensions of the critic network do not change with the number of agents. Finally, 8 RCs can help the MEMG provide 21.65 MWh of load restoration in the power network and 26.61 M S m^3 h of load restoration in the gas network, which is 92.27% and 95.39% of baseline load, respectively.

Nevertheless, the research limitations and future extensions of this work should be discussed. First, the environment state introduced in this paper focuses on the observed information at the current time step only. However, considering that RL is operating in a dynamic process and some of the state features (e.g., traffic volumes, PV power generation, power and gas demand) capture the time-series characteristics, future work will make use of the advanced forecasting method, support vector regression (SVR) [49] or long short-term memory (LSTM) [50] techniques, to predict the future trends of time-series states and feed it into the RL training process. Second, this paper only focuses on the mobile sources of RC technology. Future work will include various mobile power sources to provide resilience, such as mobile emergency generators (MEGs), mobile energy storage systems (MESSs), electric vehicles (EVs), etc. Third, this paper only solves the load restoration problem of a MEMG after an outage occurs. However, the other stages of resilience are not considered. Future work will focus on the coordination of different stages (e.g., vulnerability assessment, preventive allocation, and restoration action) towards system resilience enhancement.

CRediT authorship contribution statement

Dawei Qiu: Methodology, Software, Data curation, Validation, Formal analysis, Writing – original draft, Writing – review & editing. **Yi Wang:** Methodology, Software, Data curation, Validation, Formal analysis, Writing – original draft, Writing – review & editing. **Tingqi Zhang:** Methodology, Writing – original draft, Writing – review & editing. **Mingyang Sun:** Methodology, Writing – original draft, Writing – review & editing. **Goran Strbac:** Conceptualization, Project administration, Methodology, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

This work was supported by two UK EPSRC projects: 'Integrated Development of Low-Carbon Energy Systems (IDLES): A Whole-System Paradigm for Creating a National Strategy' (project code: EP/R045518/1) and UK-China project - 'Technology Transformation to Support Flexible and Resilient Local Energy Systems' (project code: EP/T021780/1), and one Horizon Europe project: 'Reliability, Resilience and Defense technology for the griD' (Grant agreement ID: 101075714) as well as the National Natural Science Foundation of China under Grants 62103371, 52161135201, U20A20159.

References

[1] Hussain A, Bui V-H, Kim H-M. Microgrids as a resilience resource and strategies used by microgrids for enhancing resilience. Appl Energy 2019;240:56–72.
[2] Sayed AR, Wang C, Bi T. Resilient operational strategies for power systems considering the interactions with natural gas systems. Appl Energy 2019;241:548–66.
[3] Wang Y, Rousis AO, Strbac G. On microgrids and resilience: A comprehensive review on modeling and operational strategies. Renew Sustain Energy Rev 2020;134:110313.
[4] Wu R, Sansavini G. Integrating reliability and resilience to support the transition from passive distribution grids to islanding microgrids. Appl Energy 2020;272:115254.
[5] He C, Dai C, Wu L, Liu T. Robust network hardening strategy for enhancing resilience of integrated electricity and natural gas distribution systems against natural disasters. IEEE Trans Power Syst 2018;33(5):5787–98.
[6] Lei S, Chen C, Li Y, Hou Y. Resilient disaster recovery logistics of distribution systems: Co-optimize service restoration with repair crew and mobile power source dispatch. IEEE Trans Smart Grid 2019;10(6):6187–202.
[7] Lin Y, Chen B, Wang J, Bie Z. A combined repair crew dispatch problem for resilient electric and natural gas system considering reconfiguration and DG islanding. IEEE Trans Power Syst 2019;34(4):2755–67.
[8] Mishra S, Anderson K, Miller B, Boyer K, Warren A. Microgrid resilience: A holistic approach for assessing threats, identifying vulnerabilities, and designing corresponding mitigation strategies. Appl Energy 2020;264:114726.
[9] Arif A, Wang Z, Chen C, Wang J. Repair and resource scheduling in unbalanced distribution systems using neighborhood search. IEEE Trans Smart Grid 2019;11(1):673–85.
[10] Ding T, Wang Z, Jia W, Chen B, Chen C, Shahidehpour M. Multiperiod distribution system restoration with routing repair crews, mobile electric vehicles, and soft-open-point networked microgrids. IEEE Trans Smart Grid 2020;11(6):4795–808.
[11] Ye Z, Chen C, Chen B, Wu K. Resilient service restoration for unbalanced distribution systems with distributed energy resources by leveraging mobile generators. IEEE Trans Ind Inform 2021;17(2):1386–96.
[12] Arif A, Ma S, Wang Z, Wang J, Ryan SM, Chen C. Optimizing service restoration in distribution systems with uncertain repair time and demand. IEEE Trans Power Syst 2018;33(6):6828–38.
[13] Taheri B, Safdarian A, Moeini-Aghtaie M, Lehtonen M. Distribution system resilience enhancement via mobile emergency generators. IEEE Trans Power Deliv 2020;36(4):2308–19.
[14] Li J, Khodayar ME, Feizi MR. Hybrid modeling based co-optimization of crew dispatch and distribution system restoration considering multiple uncertainties. IEEE Syst J 2021.
[15] Sayed AR, Wang C, Bi T. Resilient operational strategies for power systems considering the interactions with natural gas systems. Appl Energy 2019;241:548–66.
[16] Sang M, Ding Y, Bao M, Li S, Ye C, Fang Y. Resilience-based restoration strategy optimization for interdependent gas and power networks. Appl Energy 2021;302:117560.
[17] Bao M, Ding Y, Sang M, Li D, Shao C, Yan J. Modeling and evaluating nodal resilience of multi-energy systems under windstorms. Appl Energy 2020;270:115136.

[18] Sutton RS, Barto AG. Reinforcement learning: an introduction. MIT Press; 2018.
[19] Perera A, Kamalaruban P. Applications of reinforcement learning in energy systems. Renew Sustain Energy Rev 2021;137:110618.
[20] Ruan G, Zhong H, Zhang G, He Y, Wang X, Pu T. Review of learning-assisted power system optimization. CSEE J Power Energy Syst 2021;7(2):221–31.
[21] Li Q, Zhang X, Guo J, Shan X, Wang Z, Li Z, Chi KT. Integrating reinforcement learning and optimal power dispatch to enhance power grid resilience. IEEE Trans Circuits Syst II: Express Briefs 2021.
[22] Yao S, Gu J, Zhang H, Wang P, Liu X, Zhao T. Resilient load restoration in microgrids considering mobile energy storage fleets: A deep reinforcement learning approach. In: 2020 IEEE power energy soc gen meeting. IEEE; 2020, p. 1–5.
[23] Zhou Z-c, Wu Z, Jin T. Deep reinforcement learning framework for resilience enhancement of distribution systems under extreme weather events. Int J Elect Power Energy Syst 2021;128:106676.
[24] Hosseini MM, Parvania M. Resilient operation of distribution grids using deep reinforcement learning. IEEE Trans Ind Inform 2021;18(3):2100–9.
[25] Dehghani NL, Jeddi AB, Shafieezadeh A. Intelligent hurricane resilience enhancement of power distribution systems via deep reinforcement learning. Appl Energy 2021;285:116355.
[26] Sun J, Zhang Z. A post-disaster resource allocation framework for improving resilience of interdependent infrastructure networks. Transp Res D: Transp Environ 2020;85:102455.
[27] Zhou Q, Li Y, Zhao D, Li J, Williams H, Xu H, Yan F. Transferable representation modelling for real-time energy management of the plug-in hybrid vehicle based on k-fold fuzzy learning and Gaussian process regression. Appl Energy 2022;305:117853.
[28] Zhou Q, Zhao D, Shuai B, Li Y, Williams H, Xu H. Knowledge implementation and transfer with an adaptive learning network for real-time power management of the plug-in hybrid vehicle. IEEE Trans Neural Netw Learn Syst 2021;32(12):5298–308.
[29] Nie H, Chen Y, Xia Y, Huang S, Liu B. Optimizing the post-disaster control of islanded microgrid: A multi-agent deep reinforcement learning approach. IEEE Access 2020;8:153455–69.
[30] Wang Y, Qiu D, Strbac G. Multi-agent deep reinforcement learning for resilience-driven routing and scheduling of mobile energy storage systems. Appl Energy 2022;310:118575.
[31] Wu T, Wang J, Lu X, Du Y. AC/DC hybrid distribution network reconfiguration with microgrid formation using multi-agent soft actor-critic. Appl Energy 2022;307:118189.
[32] Kamruzzaman M, Duan J, Shi D, Benidris M. A deep reinforcement learning-based multi-agent framework to enhance power system resilience using shunt resources. IEEE Trans Power Syst 2021;36(6):5525–36.
[33] Sutton RS, Precup D, Singh S. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. Artif Intell 1999;112(1–2):181–211.
[34] Yu C, Velu A, Vinitsky E, Wang Y, Bayen A, Wu Y. The surprising effectiveness of mappo in cooperative, multi-agent games. 2021, arXiv preprint arXiv:2103.01955.
[35] Terry JK, Grammel N, Hari A, Santos L. Parameter sharing is surprisingly useful for multi-agent deep reinforcement learning. 2020, p. arXiv–2005.
[36] Wang Y, Chen C, Wang J, Baldick R. Research on resilience of power systems under natural disasters—A review. IEEE Trans Power Syst 2016;31(2):1604–13.
[37] Yuanqing W, Wei Z, Lianen L. Theory and application study of the road traffic impedance function. J Highway Transp Res Dev 2004;21(9):82–5.
[38] Yang Z, Zhong H, Bose A, Zheng T, Xia Q, Kang C. A linearized OPF model with reactive power and voltage magnitude: A pathway to improve the MW-only DC OPF. IEEE Trans Power Syst 2018;33(2):1734–45.
[39] Ding T, Hu Y, Bie Z. Multi-stage stochastic programming with nonanticipativity constraints for expansion of combined power and natural gas systems. IEEE Trans Power Syst 2018;33(1):317–28.
[40] Lei S, Chen C, Zhou H, Hou Y. Routing and scheduling of mobile power sources for distribution system resilience enhancement. IEEE Trans Smart Grid 2019;10(5):5650–62.
[41] Wang C, Wei W, Wang J, Liu F, Qiu F, Correa-Posada CM, Mei S. Robust defense strategy for gas–electric systems against malicious attacks. IEEE Trans Power Syst 2017;32(4):2953–65.
[42] Ratnam EL, Weller SR, Kellett CM, Murray AT. Residential load and rooftop PV generation: an Australian distribution network dataset. Int J Sustain Energy 2017;36(8):787–806.
[43] USFederal Highway Administration and Environmental Protection Agency (FHWA and EPA). National NearRoad study. 2022, URL https://www.fhwa.dot.gov/ENVIRonment/air_quality/air_toxics/research_and_analysis/near_road_study/protocol/protocol03.cfm.
[44] Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal policy optimization algorithms. 2017, arXiv preprint arXiv:1707.06347.
[45] Whitley D. A genetic algorithm tutorial. Stat Comput 1994;4(2):65–85.
[46] Camacho EF, Alba CB. Model predictive control. Springer science & business media; 2013.
[47] Kingma DP, Ba J. Adam: A method for stochastic optimization. In: Proc. 3rd int. conf. learn. represent. (ICLR). San Diego, USA; May. 2015, p. 1–15.
[48] Dolatabadi SH, Ghorbanian M, Siano P, Hatziargyriou ND. An enhanced IEEE 33 bus benchmark test system for distribution system studies. IEEE Trans Power Syst 2020;36(3):2565–72.
[49] Karasu S, Altan A, Bekiros S, Ahmad W. A new forecasting model with wrapper-based feature selection approach using multi-objective optimization technique for chaotic crude oil time series. Energy 2020;212:118750.
[50] Karasu S, Altan A. Crude oil time series prediction model based on LSTM network with chaotic henry gas solubility optimization. Energy 2022;242:122964.