frontiers | Frontiers in Energy Research

TYPE Original Research
PUBLISHED 01 February 2024
DOI 10.3389/fenrg.2024.1357406

OPEN ACCESS

EDITED BY
Kaiping Qu,
China University of Mining and Technology, China

REVIEWED BY
Weiyu Wang,
Changsha University of Science and Technology, China
Cao Yingping,
Hong Kong Polytechnic University, Hong Kong SAR, China

*CORRESPONDENCE
Xing-Chen Shangguan,
star@cug.edu.cn

RECEIVED 18 December 2023
ACCEPTED 15 January 2024
PUBLISHED 01 February 2024

CITATION
Gao Z, Kang W, Chen X, Gong S, Liu Z, He D, Shi S and Shangguan X-C (2024), Optimal economic dispatch of a virtual power plant based on gated recurrent unit proximal policy optimization.
Front. Energy Res. 12:1357406.
doi: 10.3389/fenrg.2024.1357406

Optimal economic dispatch of a virtual power plant based on gated recurrent unit proximal policy optimization

Zhiping Gao[1], Wenwen Kang[1], Xinghua Chen[1], Siru Gong[1], Zongxiong Liu[1], Degang He[2], Shen Shi[3] and Xing-Chen Shangguan[3]*

[1]Hubei Branch of State Power Investment Group Co, Ltd, Wuhan, China, [2]Institute of New Energy, Wuhan, China, [3]School of Automation, China University of Geosciences, Wuhan, China

The intermittent renewable energy in a virtual power plant (VPP) brings generation uncertainties, which prevents the VPP from providing a reliable and user-friendly power supply. To address this issue, this paper proposes a gated recurrent unit proximal policy optimization (GRUPPO)-based optimal VPP economic dispatch method. First, electrical generation, storage, and consumption are established to form a VPP framework by considering the accessibility of VPP state information. The optimal VPP economic dispatch can then be expressed as a partially observable Markov decision process (POMDP) problem. A novel deep reinforcement learning method called GRUPPO is further developed based on VPP time series characteristics. Finally, case studies are conducted over a 24-h period based on the actual historical data. The test results illustrate that the proposed economic dispatch can achieve a maximum operation cost reduction of 6.5% and effectively smooth the supply–demand uncertainties.

KEYWORDS

virtual power plant, demand response, deep reinforcement learning, gated recurrent unit, proximal policy optimization

1 Introduction

1.1 Background and motivation

With the global energy shortage and environmental deterioration becoming increasingly prominent, distributed renewable energy resources have gained popularity in the power system and developed rapidly (Naveen et al., 2020; Huang et al., 2021; Liu et al., 2023). Although the renewable energy implementation can generally reduce the dependence on fossil generation, the low unit capacity and high fluctuation hinder its reliable supply. As a result of inherent temporal–spatial complementarities, virtual power plants (VPPs) integrated with cooperative and transactive energy management can effectively cope with the core issues and enhance the overall economy (Koraki and Strunz, 2017).

The VPP is defined as an aggregator of distributed supply–demand resources, which would independently perform a transactive behavior with the market or operator (Etherden et al., 2015; Lin et al., 2020; Gough et al., 2022). However, due to its lower capacity and inherent sporadic nature, its integration into the current power system is complicated (Xu et al., 2021). Although VPPs have developed rapidly, the high penetration of renewable energy and the proactive end

users make VPPs more uncertain. These uncertainties cause disturbances in the optimal VPP economic dispatch and prevent VPPs from providing a reliable and user-friendly power supply. Therefore, it is essential to design an effective VPP economic dispatch method to enhance economic benefits and smooth the supply–demand uncertainties.

1.2 Literature review

In order to handle the uncertainties in the VPP dispatch, various optimization methods have been proposed, including stochastic optimization and robust optimization. Liu et al. (2018) proposed an interval-deterministic combined optimization method to maximize the deterministic profits and profit intervals of VPPs (Liu et al., 2018). A data-adaptive robust optimization method was proposed by Zhang et al. (2018) to optimize the dispatch scheme with adjustable robustness parameters. A deterministic price-based unit commitment was proposed by Mashhour and Moghaddas-Tafreshi (2010), and a genetic algorithm was used to solve the uncertainties. Chen et al. (2018) presented a fully distributed method for VPP economic dispatch using the alternating direction multiplier method (ADMM) and the consensus algorithm (Chen et al., 2018). Yang et al. (2013) proposed a consensus-based distributed economic dispatch algorithm through the iterative coordination of local agents.

When faced with the supply–demand uncertainties, these traditional optimization-based methods usually rely on accurate system models and *a priori* knowledge, which are difficult to obtain in practice (Xu et al., 2019). Although robust optimization methods can deal with uncertainties to some extent, these methods are very conservative. Meanwhile, these methods cannot deal with dynamic and random changes, due to which real-time information and interactions with various energy sources may not be able to capture. Traditional optimization-based methods also rely on reliable solvers or heuristic algorithms (Xu et al., 2020), which is time-consuming and cannot meet the real-time requirements of practical VPP problems.

Reinforcement learning has become a highly effective approach for addressing optimization problems in various domains (Książek et al., 2019). Unlike traditional optimization methods that often rely on extensive domain knowledge or problem-specific heuristics, reinforcement learning allows agents to discover effective strategies through trial-and-error processes (Bui et al., 2020). Reinforcement learning is well suited for sequential decision-making. In many optimization problems, decisions must be made in a sequential manner with each decision influencing future decisions. Reinforcement learning algorithms, such as Q-learning and policy gradient methods, explicitly model this sequential aspect of decision-making by updating the agent's policy based on the outcomes of previous actions. This allows the agent to learn optimal sequences of decisions that lead to desired outcomes (Huang et al., 2021). In many real-world optimization problems, the agent may not have complete information about the state of the system. Reinforcement learning agents learn to make decisions based on partial information, effectively reasoning about the most likely state of the system and taking actions accordingly. This ability to handle incomplete information makes reinforcement learning suitable for a wide range of real-world optimization problems with uncertainties.

Deep reinforcement learning integrates deep learning and reinforcement learning, which has been widely adopted for solving VPP problems in the Internet of Energy (IoE) domain. For instance, Sun et al. (Hua et al., 2019) mainly studied IoE management, and reinforcement learning was adopted to formulate the best operating strategies. Du et al. (2018) studied the IoE architecture design and adopted reinforcement learning to optimize electric vehicle charging. Liu et al. (2018) combined deep learning with reinforcement learning for improving the generating unit tripping strategy. Combining reinforcement learning and deep neural network, Lu et al. (2019) presented a demand response algorithm for the IoE system based on real-time execution. However, the reinforcement learning methods in the above studies are all based on Q-learning or deep Q-learning methods, which are limited to discrete action spaces. To address this problem, Zhao et al. (2022) adopted a proximal policy optimization (PPO)-based reinforcement learning method, which contains both continuous and discrete action spaces. Zhao et al. (2022) proved that the system cost is reduced by 12.17% compared to the Q-learning method.

However, two problems still remain to be addressed in the existing reinforcement learning-based VPP economic dispatch method. The first problem is that the historical VPP information is not considered in the above studies. Actually, the VPP economic dispatch cannot follow the Markov decision process (MDP) since the integration of renewable energy sources makes it a sequential decision process problem. As a type of artificial neural network, the recurrent neural network (RNN) is commonly used to address these ordinal or temporal problems, which can extract the time series information effectively. The gated recurrent unit (GRU), which optimizes the update and reset gates, is another type of the long short-term memory network. Compared to RNN, GRU offers computational efficiency, superior long-term dependency capture, effective vanishing gradient solution, and remarkable generalization capabilities (Canizo et al., 2019). These key advantages make the GRU an excellent choice for various sequential learning tasks, particularly in domains where capturing long-term dependencies is of paramount importance (Thanh et al., 2022).

The second problem is that the above methods all need a central agent to coordinate the VPP supply–demand balance. Actually, the VPP would not be managed with a single operator, and this centralized management would give rise to various disadvantages, including intensive information transmission and low-efficiency operation. It is a foreseeable trend that the VPP would gradually form a distributed manner, which could potentially satisfy geographical end users. Various decomposition techniques, including ADMM (Chen et al., 2018; Xu et al., 2019) and consensus algorithm (Yang et al., 2013), have been successfully applied for decentralized/distributed decision-making. Compared with the central method based on the single agent, the multi-agent optimization method can assign dispatch tasks to multiple agents for processing, which improves the processing capacities and solution efficiency. In addition, even if one agent fails or another agent is added, the entire system can still maintain a stable operation. In other words, the multi-agent approach will be more scalable, adaptive, and robust (Gronauer et al., 2023).

1.3 Contribution

To sum up the above discussion, this paper proposes a gated recurrent unit proximal policy optimization (GRUPPO)-based

optimal VPP economic dispatch method. The contributions of this article are as follows:

(1) The PPO-based deep reinforcement learning method is developed to handle both continuous and discrete action spaces. Compared with the traditional method, including the deterministic optimization and robust optimization in Mashhour and Moghaddas-Tafreshi (2010), Yang et al. (2013), Chen et al. (2018), Liu et al. (2018), and Zhang et al. (2018), the proposed approach can better deal with the supply–demand uncertainties and meet the real-time economic dispatch requirement for the VPP.
(2) The GRU network is equipped into the PPO-based deep reinforcement learning method to form the GRUPPO approach. Different from the reinforcement learning approaches in Sun et al. (2017), Du et al. (2018), Liu et al. (2018), and Lu et al. (2019), the proposed GRUPPO scheme can fully consider the historical time series information for economic decision-making, effectively reducing the VPP operation cost.
(3) A multi-agent optimization framework is developed to capture the distributed characteristics in the VPP. The optimization framework adopts centralized training and distributed execution, thereby performing higher flexibility and scalability against complex situations.

The remainder of this paper is organized as follows: in Section 2, the modeling of the VPP economic dispatch is established, and its objective function is designed. In Section 3, the GRUPPO strategy and its multi-agent framework are proposed for the optimal VPP economic dispatch. In Section 4, case studies are conducted based on the actual historical data. Conclusions are drawn in Section 5.

2 VPP economic dispatch

2.1 Framework and assumptions

The VPP leverages advanced information communication technology to aggregate and coordinate multiple distributed energy resources. The core concept of a VPP is aggregation and coordination. The following assumptions and simplifications are considered:

1) The VPP is assumed to have access to real-time data on generation, demand, and grid conditions. These data are necessary for the VPP to make decisions about power generation and distribution.
2) The VPP is assumed to have efficient and reliable control and communication systems to coordinate multiple distributed energy resources.
3) Thermal properties of heating, ventilation, and air-conditioning (HVAC), including heat generation, storage, and transfer, are assumed to happen only in thermal nodes.

Though raising concerns over the inaccuracy issues, reasonable assumptions and simplifications here could render the model a more computationally tractable and more practically meaningful analysis.

2.2 VPP supply–demand model

The VPP components comprise thermal power generation, photovoltaic generation, battery energy storage, the basic load, the power flexible load, and the temperature-adjustable load.

1) Thermal power generation unit

The VPP relies on small-scale thermal power units to maintain the flexibility and stability. The operation of thermal power generation unit in the VPP meets the output constraints and the ramp constraints:

$$P_{TH,min} \le P_{TH,t} \le P_{TH,max}; \tag{1}$$

$$R_{TH,min} \le P_{TH,t} - P_{TH,t-1} \le R_{TH,max}, \tag{2}$$

where $P_{TH,t}$ and $P_{TH,t-1}$ are the thermal power output at moments t and $t-1$, respectively; $P_{TH,min}$, $P_{TH,max}$, $R_{TH,min}$, and $R_{TH,max}$ are minimum output power, maximum output power, ramp-down power, and ramp-up power of the thermal unit, respectively.

2) Power flexible loads

The power flexible loads, including LED lights with adjustable brightness or electric fans with adjustable speed, can participate in the VPP economic dispatch as a flexible load. In general, these loads can be adjusted within the rated capacity range. Their total power needs to meet the following constraints:

$$P_{pf,min} \le P_{pf,t} \le P_{pf,max}; \tag{3}$$

$$P_{pf,exp,t} = P_{pf,exp,t-1} + \left(P_{pf,t}\right); \tag{4}$$

$$P_{pf,exp,T} \ge P_{pf,exp}, \tag{5}$$

where $P_{pf,min}$ and $P_{pf,max}$ are minimum power and maximum power of these loads, respectively; $P_{pf,exp,t}$, $P_{pf,exp,T}$, and $P_{pf,exp}$ are total power of the previous moment t, the total power of the whole time period T, and the minimum power to meet user needs, respectively.

3) Temperature flexible loads

Heating loads, including pitch heating, water heating, and HVAC, are taken as temperature-adjustable loads. The common feature of these heating loads is that the operating temperature t can be adjusted according to artificial settings. The working temperature t should be enforced to ensure the safe and reliable operation of the equipment:

$$T_{HVAC,t} = T_{HVAC,t-1} + a_1\left(T_{out,t} - T_{HVAC,t-1}\right) + a_2 a_{HVAC} P_{HVAC,heat,t} - a_2\left(1 - a_{HAVC}\right) P_{HVAC,cool,t}; \tag{6}$$

$$a_{HVAC} \in \{0, 1\}; \tag{7}$$

$$T_{min} \le T_{HVAC,t} \le T_{max}, \tag{8}$$

where $T_{\min}$ and $T_{\max}$ are the lower and upper temperatures, respectively; a_1 and a_2 are the physical parameters which is jointly calculated via thermal capacities and resistances; $T_{HVAC,t-1}$ is the temperature of the last time moment $t-1$; $T_{out,t}$ represents the

current outside temperature; $P_{HVAC,heat,t}$ and $P_{HVAC,cool,t}$ denote the heating and cooling power of the air-conditioner, respectively; and a_{HVAC} is the running state of the air-conditioner, where a_{HVAC} = 1 and 0 indicate that the air-conditioner is in a state of heating and cooling, respectively.

4) Battery energy storage

Battery energy storage is an important energy storage, which has the advantages of strong environmental adaptability, short construction period, and convenient small-scale configuration. The charge and discharge states of the battery energy storage system must be limited to a certain range so as to avoid overcharge or discharge:

$$SOC_t = SOC_{t-1} + \frac{\eta_{ch} a_{SOC,t} P_{SOC,ch,t-1} \Delta t}{E_{SOC}} - \frac{(1 - a_{SOC,t}) P_{SOC,dis,t-1} \Delta t}{\eta_{dis} E_{SOC}}; \tag{9}$$

$$a_{SOC,t} \in \{0, 1\}, \tag{10}$$

where SOC_t and SOC_{t-1} are the current charge of the battery energy storage and the charge of the last time $t-1$, respectively; η_{ch}, η_{dis}, $P_{SOC,ch,t-1}$, and $P_{SOC,dis,t-1}$ are charge efficiency, discharge efficiency, charge power, and discharge power, respectively; Δt is the unit time; and $a_{SOC,t}$ is the status of the battery charge and discharge, where $a_{SOC,t} = 1$ indicates that the battery is being charged and $a_{SOC,t} = 0$ indicates that the battery is being discharged.

In addition, the operation of the battery energy storage needs to meet the battery capacity limit:

$$SOC_{min} \le SOC_t \le SOC_{max}; \tag{11}$$

$$0 \le P_{SOC,ch,t} \le P_{SOC,ch,max}; \tag{12}$$

$$0 \le P_{SOC,dis,t} \le P_{SOC,dis,max}, \tag{13}$$

where SOC_{min} and SOC_{max} represent the minimum and maximum battery capacities, respectively; $P_{SOC,ch,max}$ and $P_{SOC,dis,max}$ are the maximum charging and discharge power, respectively.

5) Power balance

With the regulation from energy storage and market buying/selling, power generation can be used to fulfill the power demand:

$$P_{buy,t} + P_{TH,t} + P_{SOC,dis,t} + P_{PV,t} = P_{SOC,ch,t} + P_{pf,t} + P_{HVAC,heat,t} + P_{HVAC,cool,t} + P_{sell,t}, \tag{14}$$

where $P_{buy,t}$, $P_{PV,t}$, and $P_{sell,t}$ are the purchased electricity power, photovoltaic power, and electricity sold, respectively.

2.3 Objective function

The objective function in the VPP economic dispatch consists of the coal consumption cost, battery degradation cost, air-conditioning discomfort cost, and buying/selling electricity cost.

1) Coal consumption cost

The coal consumption cost function of the thermal power unit can use the quadratic function related to the unit output:

$$C_{TH} = \sum_{t=1}^{T} a \cdot P_{TH,t}^2 + b \cdot P_{TH,t} + c, \tag{15}$$

where a, b, and c are the coefficients of the quadratic function; $P_{TH,t}$ is the power of thermal power. Through the linearization of the quadratic function, the cost function of the coal consumption is divided into M parts and denoted by

$$C_{TH} = \sum_{t=1}^{T} \sum_{m=1}^{M} K_{m,t} P_{TH,m,t} + C_t; \tag{16}$$

$$\begin{cases} C_t = a \cdot P_{TH,min}^2 + b \cdot P_{TH,min} + c; \\ 0 \le P_{TH,m,t} \le \dfrac{P_{TH,max} - P_{TH,min}}{M}; \\ P_{TH,t} = \sum_{m=1}^{M} P_{TH,m,t} + P_{TH,min}; \\ K_m = 2a(2m-1)\dfrac{P_{TH,max} - P_{TH,min}}{M} + b, \end{cases} \tag{17}$$

where $K_{m,t}$ is the slope of section m at time t of the coal consumption function after piecewise linearization; C_t is the coal consumption caused by starting up the thermal power unit and running at the minimum output $P_{TH,min}$; and $P_{TH,m,t}$ represents the output power of the thermal power unit in the m section at the t period.

2) Battery degradation cost

The battery degradation cost can be represented by

$$C_{SOC} = \sum_{t=1}^{T} \mu_{SOC} (P_{SOC,ch,t} + P_{SOC,dis,t}) \Delta t, \tag{18}$$

where T, Δt, $P_{SOC,ch,t}$, and $P_{SOC,dis,t}$ are dispatch cycle, unit time, charging power, and discharge power, respectively; μ_{soc} is the unit average/amortized degradation cost of charging/discharging over the whole service time, which can be calculated with its capital cost, cycling numbers, capacity, and reference state of charge (Xu et al., 2021).

3) Air-conditioning discomfort cost

While the constraints (Eqs 6–8) enforce the physical operation of air-conditioning, the discomfort level is introduced to measure the degree of satisfaction. The air-conditioning discomfort cost is related to the set temperature and current temperature.

$$C_{HVAC} = \sum_{t=1}^{T} \mu_{HVAC} (T_{set} - T_{HVAC,t})^2 \Delta t, \tag{19}$$

where T_{set} and $T_{HVAC,t}$ are set temperature and current time period temperature, respectively; μ_{HVAC} is the discomfort cost coefficient, which is used to measure the discomfort level.

4) Buying and selling electricity costs

The buying and selling electricity costs are calculated as follows:

$$C_{buy} = \sum_{t=1}^{T}\left(a_{buy,t}\lambda_{buy,t}P_{buy,t} - \left(1 - a_{buy,t}\right)\lambda_{sell,t}P_{sell,t}\right)\Delta t; \quad (20)$$

$$a_{buy} \in \{0, 1\}, \quad (21)$$

where $a_{buy,t}$ denotes the status of buying and selling electricity in the VPP, where $a_{buy} = 1$ means that the VPP buys electricity from the market and $a_{buy} = 0$ means that the VPP sells electricity to the market; $\lambda_{buy,t}$ and $\lambda_{sell,t}$ are electricity buying price and electricity selling price, respectively.

The objective function is defined as

$$C = \lambda_{TH}C_{TH} + \lambda_{SOC}C_{SOC} + \lambda_{HVAC}C_{HVAC} + \lambda_{buy}C_{buy}, \quad (22)$$

where λ_{TH}, λ_{SOC}, λ_{HVAC}, and λ_{buy} represent the cost coefficients of the coal consumption, battery degradation, air-conditioning discomfort, and buying and selling electricity, respectively.

3 The GRUPPO-based optimal VPP economic dispatch

In this section, the designed GRUPPO-based optimal VPP economic dispatch will be presented. First, the VPP economic dispatch is expressed as a partially observable Markov decision process (POMDP). Then, a GRUPPO-based deep reinforcement learning approach is introduced to optimize the VPP economic dispatch.

3.1 POMDP for the VPP economic dispatch

When using the reinforcement learning method to solve problems, MDP is usually used to describe the environment. MDP is characterized by the environment that is completely observable, and the current state can fully represent the process. That is, according to the current state, the next state can be deduced, the current state captures all relevant information from history, and the current state is a sufficient statistic for the future. However, for the VPP economic dispatch problem, the model contains random renewable energy. In the dispatch process, the next state of the VPP is not only completely determined by the current state but also depends on external random factors. The model state is not completely observable, and it is reasonable to express the VPP economic dispatch problem as a POMDP. Its structure diagram is shown in Figure 1. Generally, POMDP can be realized as a 7-tuple model $\{S, A, s, a, T, R, \lambda\}$ (Wang et al., 2023).

The VPP model shown in Figure 1 represents the environment, and the agent is a hypothetical entity responsible for the VPP economic dispatch. The agent makes a corresponding decision based on the state of the environment, where the state and decision represent the observations and actions of the agent, respectively. The environment accepts the action of the agent and produces the corresponding change, which depends on the state transfer function $T(s_t, a_t, \chi)$. The environment gives the corresponding reward according to the agent action, and the reward received by the agent is related to the objective function of the VPP economic dispatch.

1) Environment

Considering the supply–demand uncertainties, the reinforcement learning environment operates according to the individual device models in Chapter 2 and also needs to satisfy their physical constraints in Chapter 2. These devices include thermal power generation, photovoltaic power generation, battery storage, base load, flexible load, and temperature-adjustable load.

2) Agent

The VPP dispatch agent is a deep neural network, which obtains the reward by constantly interacting with the environment and then updates the neural network parameters according to the reward. The interaction process between the agent and environment is to output the VPP dispatch instructions through the neural network and calculate the corresponding objective function value. The construction process of the specific agent will be described in detail in the next section.

3) State and observation

The agent needs to implement the corresponding action according to the environment state, which is the state space. For the VPP economic dispatch, the state observation space $s_t \in S$ of the agent is shown as follows:

$$s_t = \left\{\lambda_{buy,t}, \lambda_{sell,t}, P_{TH,t}, SOC_t, T_{HVAC,t}, P_{PV,t}, P_{base,t}, P_{pf,exp,t}\right\}. \quad (23)$$

4) Action space

The action carried out by the agent according to the environment state is the action space. Lower-dimensional actions help the agent learn faster. Since battery charging and discharging cannot take place simultaneously, the charging and discharging of the battery are combined into a single action (instead of positive and negative). The same applies to the air-conditioner. For the economic dispatch task, the action space is expressed as follows:

$$a_t = \left\{P_{TH,t}, P_{HVAC,t}, P_{SOC,t}, P_{pf,t}\right\}; \quad (24)$$

$$P_{SOC,t} = a_{SOC,t}P_{SOC,ch,t} - \left(1 - a_{SOC,t}\right)P_{SOC,dis,t}; \quad (25)$$

$$P_{HVAC,t} = a_{HVAC,t}P_{HVAC,heat,t} - \left(1 - a_{HVAC,t}\right)P_{HVAC,cool,t}. \quad (26)$$

5) State transition

Based on a policy $\pi(a_t|s_t)$, the agent could calculate and perform an action after its current observation s_t. Afterward, based on the state transition function $s_{t+1} = T(s_t, a_t, \chi)$, the environment proceeds to s_{t+1}, which is impacted by state s_t, actions a_t, and the environmental randomness χ_t.

Here, $\chi_t = [\lambda_{buy,t}, \lambda_{sell,t}, T_{HVAC,t}, P_{PV,t}, P_{base,t}]$ indicates the exogenous states, which are unrelated to the agent's actions and show model variability. In general, reinforcement learning could cope with such variabilities in a data-driven way. It does not rely on precise probability uncertainty distributions and updates state

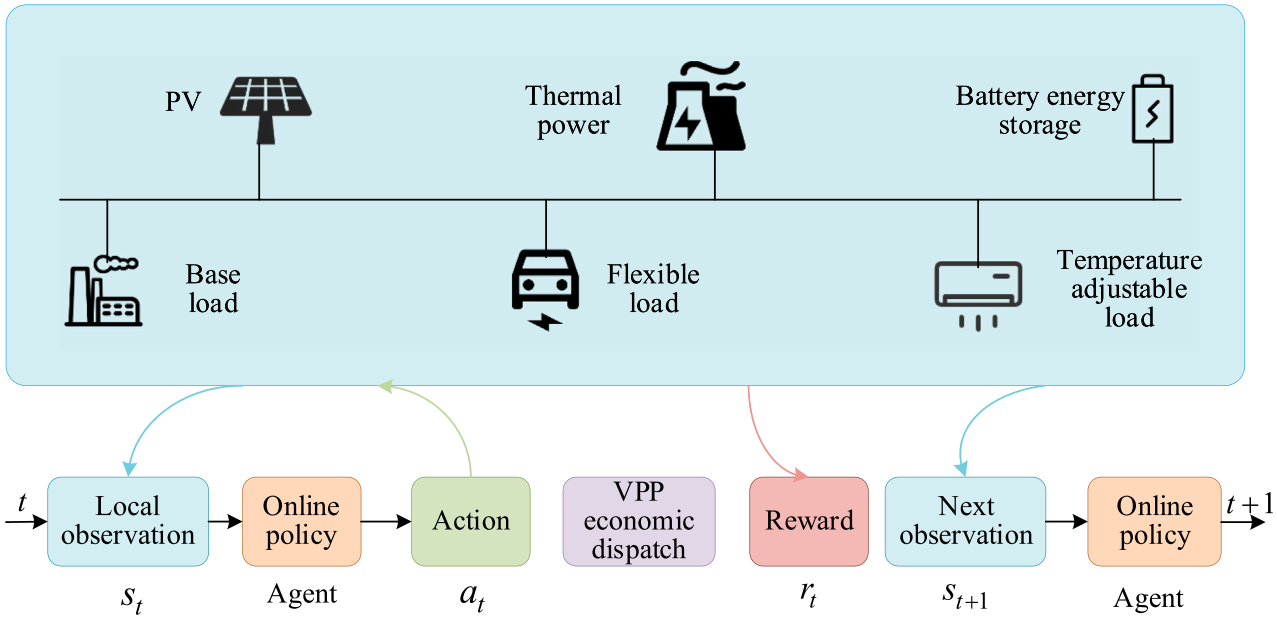

FIGURE 1
POMDP of the VPP economic dispatch.

characteristics from the dataset. The state of $\chi'_t = [P_{TH,t}, SOC_t, T_{HVAC,t}, P_{pf,exp,t}]$ has no association with the external environment but is associated only with the policy $\pi(a_t|s_t)$. The state update is required to satisfy the system constraints.

6) Reward

The reward function is to drive the agent's decision-making, and reward signals can be of any value (Canizo et al., 2019). The reward function is generally set in the range of 0–1 to enhance and ensure the convergence and optimality. Since the goal is to minimize the dispatch cost, the establishment function of each time step is designed as follows:

$$r_t = \lambda_{TH} C_{TH,t} + \lambda_{SOC} C_{SOC,t} + \lambda_{HVAC} C_{HVAC,t} + \lambda_{buy} C_{buy,t}. \quad (27)$$

7) Objective

Each episode is divided into discrete time nodes $t \in \{0,1,2, \ldots, T\}$. The agent starts from an initial state s_0. At each time point t, the agent moves to the next state s_{t+1} based on the observation of the environment state s_t, action a_t, and an immediate reward r_t. Based on this, the agent creates its trajectories of observations, actions, and rewards: $\tau = s_0, a_0, r_0, s_1, a_1, r_1 \ldots, r_T$. In the POMDP, the agent seeks an optimal policy $\pi(a_t|s_t)$ for the maximization of the discounted reward:

$$R = \sum_{t=0}^{T} \gamma^t r_t, \quad (28)$$

where $\gamma \in [0,1]$ is the discount factor to decide the importance of immediate and future rewards.

3.2 GRUPPO-based deep reinforcement learning

In this subsection, a reinforcement learning method called GRUPPO is used for optimizing the VPP economic dispatch based on the POMDP. The GRUPPO approach includes the following three crucial steps:

1) Update the dispatch policy via a standard PPO algorithm

PPO, as a policy gradient algorithm, has been employed in a multitude of optimization models. Generally, PPO is featured by an actor–critic network and is able to handle high-dimensional continuous spaces. Through the Gaussian distribution, a stochastic policy $\pi_\theta(a_t|s_t)$ of the actor network could be developed to feature the continuous action spaces in (24). It gives the standard deviation σ and mean μ, sampling the action a_t on s_t for all VPP economic dispatch agents. The PPO renews the policy $\pi_\theta(a_t|s_t)$, maximizing the following clipped surrogate.

$$J_t(\theta) = \mathbb{E}_t\left[\min\left(\zeta_t \hat{A}_t, clip(\zeta_t, 1-\varepsilon, 1+\varepsilon)\hat{A}_t\right)\right], \quad (29)$$

where the product of ζ_t and $\hat{A}_t$ is the policy gradient; ζ_t is the probability ratio clipped by $clip(.)$. $\varepsilon \in [0,1]$ is used to limit the policy gradient update against its old version, if ζ_t is beyond $[1-\varepsilon, 1+\varepsilon]$. This technique ensures that the policy gradient is updated to a stable area.

ζ_t in the PPO clipped policy (29) is expressed as follows:

$$\zeta_t = \frac{\pi_\theta(a_t \mid s_t)}{\pi_\theta^{old}(a_t \mid s_t)}, \quad (30)$$

where $\pi_\theta(a_t|s_t)$ and $\pi_\theta^{old}(a_t|s_t)$ are the current and old policies, respectively. The advantage function $\hat{A}_t$ is expressed as follows:

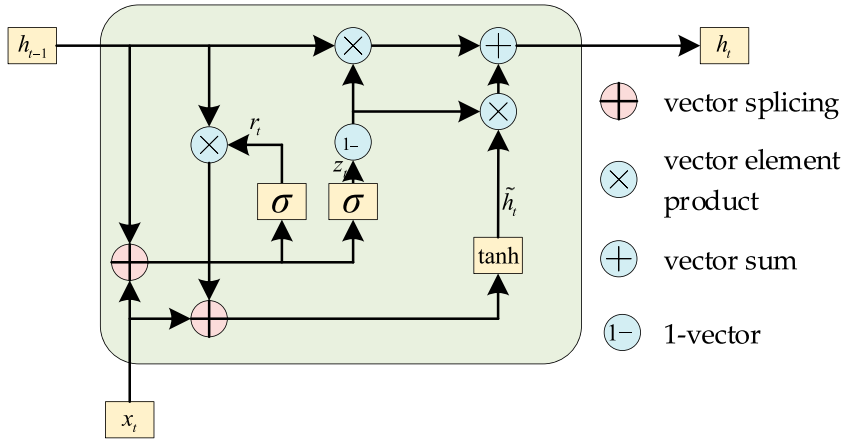

FIGURE 2
GRU network.

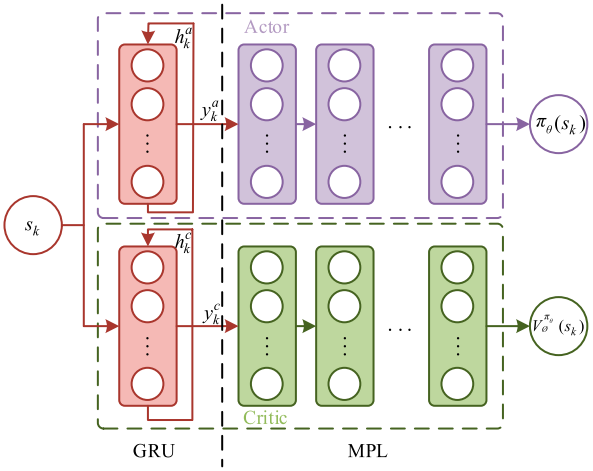

FIGURE 3
Actor–critic network.

$$\hat{A}_t = \delta_t + (\gamma\lambda)\delta_{t+1} + \cdots + (\gamma\lambda)^{T-t+1}\delta_{T-1}; \quad (31)$$

$$\delta_t = r_t + \gamma V_\phi(s_{t+1}) - V_\phi(s_{t+1}), \quad (32)$$

where $V_\phi(s)$ denotes the state-value function approximated by a critic network parameterized by ϕ; $\gamma \in [0,1]$ and $\lambda \in [0,1]$.

2) Introduce GRU into PPO to consider the time characteristics

Since the proposed VPP economic dispatch model is partially observable, the Markov property is not valid. Compared with MDP, the next state in a POMDP is not completely determined by the present observations and actions (Ma et al., 2023). Conversely, the complete history of the observation sequences ought to be taken into account. By adding a GRU layer before the multi-layer perceptron (MLP) to concisely capture the history, recursion is introduced to deal with the non-Markovian nature of the POMDP.

GRUs are well suited for capturing and modeling time series characteristics due to their ability to control information flow using gates and adapt to long-term dependencies. GRUs comprise two gates: the reset gate (r) and the update gate (z). The reset gate controls how much of the previous hidden state is passed on to the next time step, while the update gate determines how much of the new input information is incorporated into the updated hidden state. The reset gate effectively "forgets" or disregards part of the previous state, allowing the network to focus on relevant information and adapt to changing patterns in the time series. By using gates to control the flow of information, GRUs can handle long sequences more effectively than traditional RNNs. They are less likely to suffer from exploding or vanishing gradients, which can be a problem for long sequences. GRUs also have fewer parameters than some other RNN variants, making them more efficient and less prone to overfitting. When applied to a time series analysis, GRUs can capture dependencies across time steps and generate meaningful representations of the sequence data. GRUs can also be combined with other techniques, such as attention mechanisms, to further improve their performance in specific tasks.

The GRU and actor–critic networks are given in Figures 2, 3.

$$r_t = \sigma(W_r x_t, U_r h_{t-1}); \quad (33)$$

$$z_t = \sigma(W_z x_t, U_z h_{t-1}); \quad (34)$$

$$\tilde{h}_t = \tanh(W_c x_t, U_c(r_t \cdot h_{t-1})); \quad (35)$$

$$h_t = z_t \cdot h_{t-1} + (1 - z_t) \cdot \tilde{h}_t, \quad (36)$$

where r_t is the reset gate; z_t is the update gate; W_r, U_r, W_z, U_z, W_c, and U_c are neural network weight matrices; σ is the sigmoid activation function; and tanh is the hyperbolic tangent activation function.

The leveraged GRUPPO method is used to apply the PPO algorithm together with the recurrent neural network. The actor and critic networks include GRU and MLP layers. The network structure is regulated via tuning the amounts of network layers and neurons. For the activation function, tanh is chosen in the GRU layer and the output layer of MLP. In other layers of MLP, ReLU is used due to its fast convergence and low computational complexity. However, the phenomenon of gradient disappearance and gradient explosion will occur when the ReLU is used directly in the experiment. In order to solve the problem of gradient

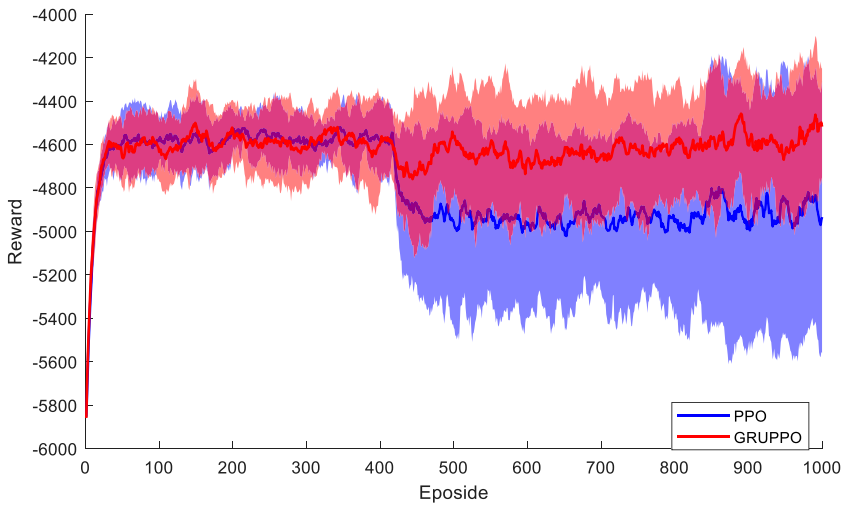

FIGURE 4
Performance comparison of PPO and GRUPPO.

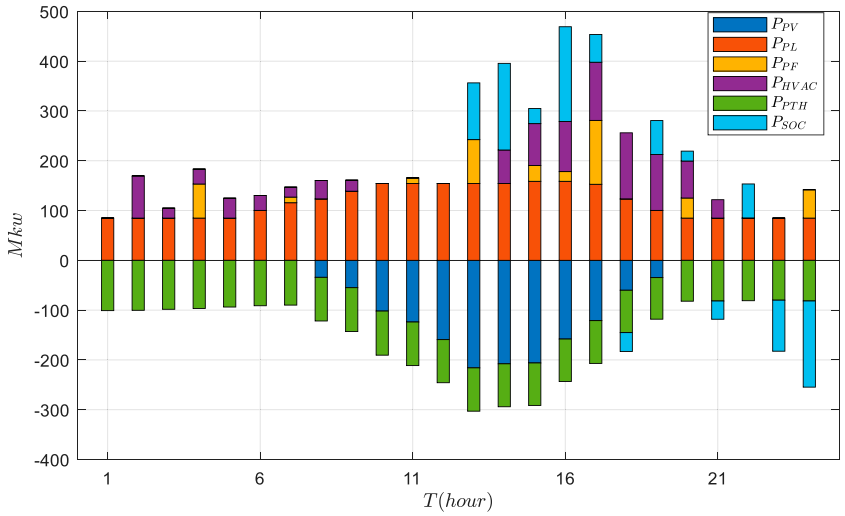

FIGURE 5
Results of the economic dispatch of the VPP under the GRUPPO approach.

disappearance and gradient explosion caused by the ReLU activation function, the layer is standardized to the neural network. In the actual experiment, this operation can effectively alleviate the phenomenon of gradient disappearance and gradient explosion so that the neural network can be trained normally.

The VPP dispatch problem using the PPO algorithm is realized through the neural network. The actor network uses the Gaussian strategy to output mean and variance. π_θ obeys the following Gaussian distribution:

$$\pi_\theta(a \mid s) = \frac{1}{\sqrt{2\pi}\sigma_\theta(s)} e^{\frac{-\left(a-\mu_\theta(s)\right)^2}{2\sigma_\theta(s)^2}}, \tag{37}$$

where a represents the action taken in state s; θ represents the policy function parameters; $\mu_\theta(s)$ represents the average value of action a in state s; and $\sigma_\theta(s)$ represents the standard deviation of action a in state s. The real action is randomly sampled according to the mean and standard deviation of the actor network output. The other network is the critic network, which outputs the value of the state according to the current state of the VPP.

3) Construct the safety layer to meet the VPP model constraints

The training reinforcement learning algorithm is an unconstrained optimization algorithm via deep neural networks, which disregards model constraints. Deploying the reinforcement learning actions to the VPP would violate the constraints, and thus a safety layer is introduced. It shows that the calculated reinforcement learning actions would be slightly updated (only when facing system safety).

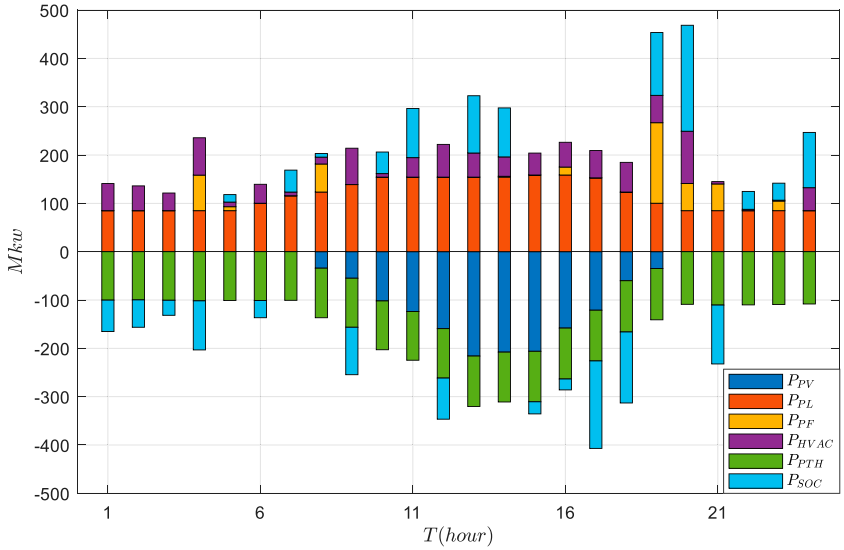

FIGURE 6
Results of the economic dispatch of the VPP under the PPO approach.

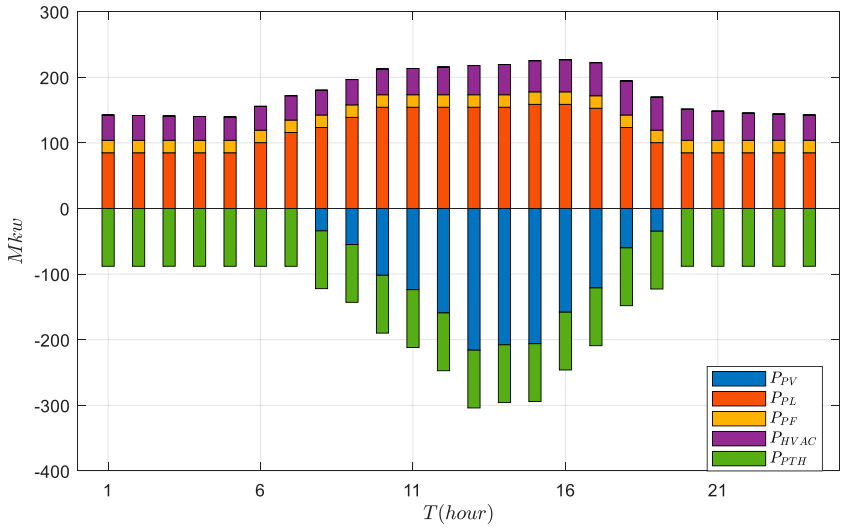

FIGURE 7
Results of the non-economic dispatch.

Safe operation is the premise of VPP economic dispatch tasks. For the VPP economic dispatch, the physical constraints are to ensure the normal operation, while the illegal actions will violate the constraints. In general, there are two ways to ensure the constraint satisfaction of the action output. The first method is to add the penalty terms for the constraint violations to the reward so that the agent can avoid making illegal action. The other method is to set the agent to take action within the allowable range. The first step is to calculate the boundary between the current state and the constraint. The action lower boundary a^- and upper boundary a^+ are calculated based on the current state s_t and constraints as follows:

$$a_t^- = \left\{P_{TH,t}^-, P_{HVAC,t}^-, P_{SOC,t}^-, P_{pf,t}^-\right\}; \tag{38}$$

$$a_t^+ = \left\{P_{TH,t}^+, P_{HVAC,t}^+, P_{SOC,t}^+, P_{pf,t}^+\right\}. \tag{39}$$

The second step is to cut the action according to the clip function. a is a constant value when the action meets the above range. When action a exceeds the boundary, the clip function is used to limit the action a within its boundary.

$$a_t = clip\left(a_t, a_t^-, a_t^+\right). \tag{40}$$

For the GRUPPO training process, the agent is equipped with $\pi_\theta(a|s)$ to interact with the environment. Then, the trajectory τ is collected and utilized to evaluate the discounted reward $\hat{R}_t = \sum_{h=t}^{T} \gamma^{h-t} r_h$. The goal of $\pi_\theta(a|s)$ is to find actions that are

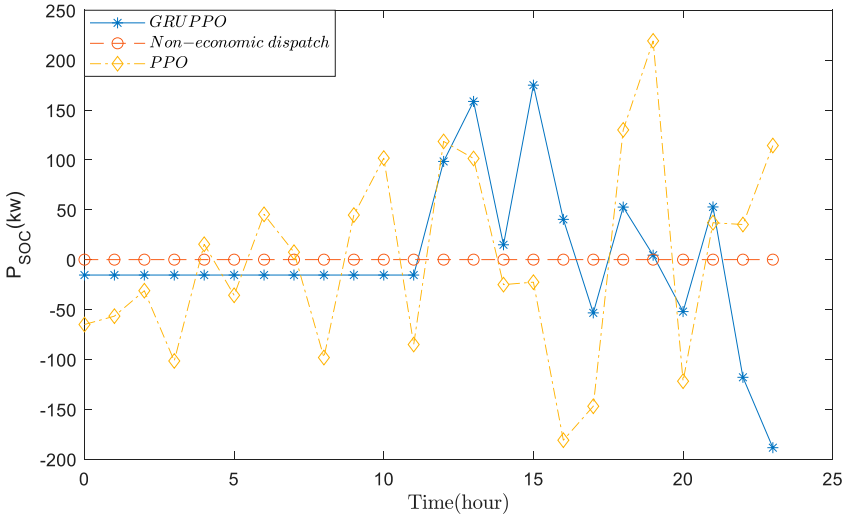

FIGURE 8
Operational results of the battery energy storage.

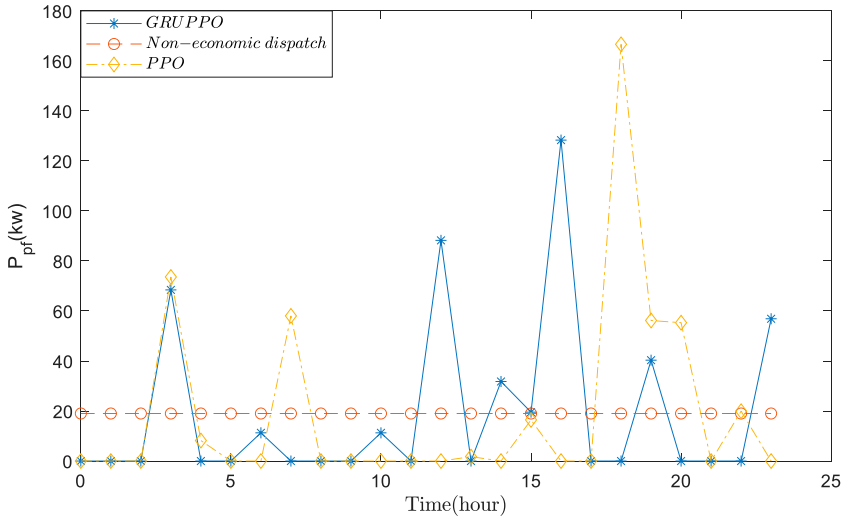

FIGURE 9
Operational results of the flexible adjustable load power.

potentially more rewarding so that they correspond to greater probabilities, and thus the strategy is more probable to choose them. For this purpose, the maximization objective function can be defined as follows:

$$\max_{\theta} J^{\alpha}(\theta) = \max_{\theta} E_{\tau\sim\pi_{\theta}} R(\tau) = \max_{\theta} \sum_{\tau} P(\tau;\theta) R(\tau). \tag{41}$$

According to the PPO algorithm, the corresponding gradient formula can be derived. $\hat{A}_t$ can also be computed with the state-value function $V_{\phi}(s)$ and trajectory τ. The actor network can be trained while maximizing

$$J^{a}(\theta) = \sum_{t=1}^{T} \min\left(\zeta_t \hat{A}_t, clip\left(\zeta_t, 1-\varepsilon, 1+\varepsilon\right)\hat{A}_t\right). \tag{42}$$

Accordingly, the critic network of GRUPPO can be trained by minimizing the following loss function of the mean-squared error:

$$J^{c}(\phi) = \sum_{t=1}^{T} \left(V_{\phi}(s_t) - \hat{R}_t\right)^2. \tag{43}$$

A weighting update for the actor and critic networks is

$$\theta \leftarrow \theta + \alpha^{\theta} \nabla_{\theta} J^{a}(\theta); \tag{44}$$

$$\phi \leftarrow \phi + \alpha^{\phi} \nabla_{\phi} J^{c}(\phi), \tag{45}$$

where α^{θ} and α^{ϕ} are the learning rates of actor and critic networks, respectively.

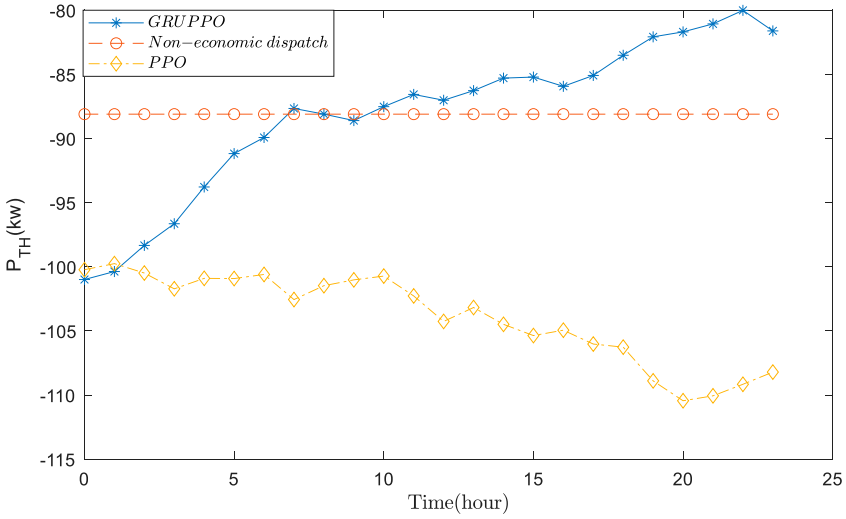

FIGURE 10
Operational results of the thermal power unit.

Finally, the pseudo-code of GRUPPO is given in Algorithm 1. First of all, GRUPPO initializes the agent's policy network and value function network. The agent collects empirical data by interacting with the environment and stores these data for subsequent training. Then, the value function network is used to compute the advantage function of the agent. Finally, the agent's strategy network is updated until a satisfactory level of performance is reached.

Initialize θ, ϕ for the actor−critic network
Set learning rates α^{θ}, α^{ϕ}
For episode (i.e., an operating day) $data = 1$ to E
Initialize VPP state s_{θ}
For the VPP agent, create a new trajectory $\tau = []$
For each time step (e.g., 1 hour) $t = 1$ to T
Chooses PPO action a_t according to observation s_t via the policy $\pi_{\theta}(a|s)$
Correct action a_t values based on the security layer
Observes reward ${r_t}^s$ and the next observation s_{t+1}
Stores the sample experience into trajectory $\tau += [s_t, {a_t}^s, {r_t}^s]$
Updates observation $s_t \rightarrow s_{t+1}$ for the VPP agent
End for
Approximates discounted reward-to-go $\hat{r}_t$ r and advantage function $\hat{A}_t$ utilizing trajectory τ
Updates the parameters θ, ϕ of networks in (44)−(45)
End for

Algorithm 1. GRUPPO for the agent.

3.3 Multi-agent framework for GRUPPO

When using a single agent for the VPP economic dispatch, the stable operation can be drastically affected by agent failure or a new plug-and-play framework. The motivation behind the multi-agent framework is to harness the power of autonomous agents and enable collaborative problem-solving in VPP systems. By distributing tasks among multiple agents, the multi-agent framework enhances scalability, robustness, adaptability, and coordination. They allow for parallel processing, fault-tolerance, and efficient utilization of resources, making them suitable for various domains and dynamic environments. The multi-agent-based GRUPPO strategy can be developed based on the above GRUPPO approach. In the multi-agent GRUPPO method, each agent is directly responsible for its own device or area, which makes it easy to expand. Here, the detailed implementation method of the multi-agent GRUPPO is given as follows:

$$s_t \in S, s_{i,t} \in s_t, \tag{46}$$

where s_t represents the overall observation value of the agent at time t; $s_{i,t}$ represents the observation value of the i agent at time t.

The training steps for the multi-agent framework differ from those of the single-agent framework, specifically in the computation of gradients and rewards. The reward function needs to compute the overall value since multiple agents are included. During training, the actor and critic network update of each agent i is

$$J^a(\theta_i) = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\min\left(\zeta_{i,t}\hat{A}_{i,t}, clip\left(\zeta_{i,t}, 1-\varepsilon, 1+\varepsilon\right)\hat{A}_{i,t}\right); \tag{47}$$

$$J^c(\phi_i) = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(V_{\phi}(s_{i,t}) - \hat{R}_t\right)^2; \tag{48}$$

$$\theta_i \leftarrow \theta_i + \alpha^{\theta_i}\nabla_{\theta_i}J^a(\theta_i); \tag{49}$$

$$\phi_i \leftarrow \phi_i + \alpha^{\phi_i}\nabla_{\phi_i}J^c(\phi_i), \tag{50}$$

where ζ_t, $\hat{A}_t$, and $V_{\phi}(s)$ represent probability ratio, advantage function, and state-value function, respectively; α^{θ} and α^{ϕ} denote the learning rates of actor and critic networks of the ith device, respectively.

Finally, the pseudo-code of multi-agent GRUPPO is given in Algorithm 2. First of all, multi-agent GRUPPO initializes the

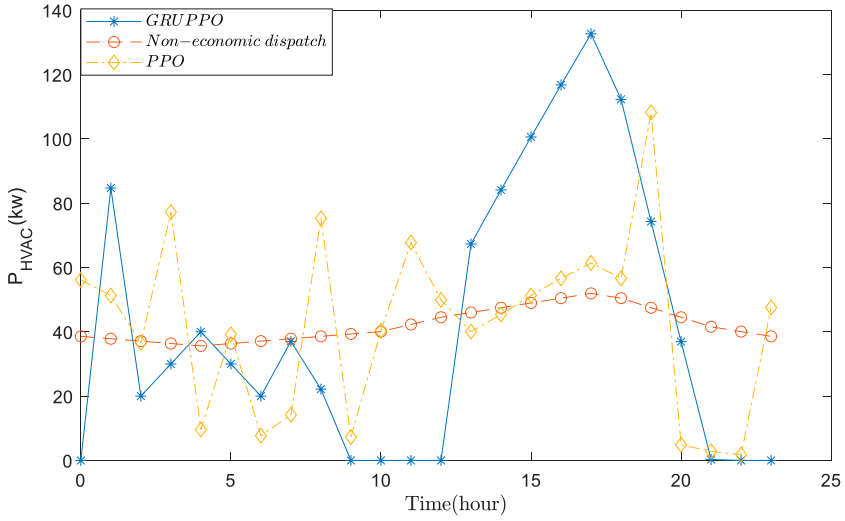

FIGURE 11
Operational results of the HVAC power.

policy network and the value function network of each agent. The agents collect experience data by interacting with the environment and store these data in a shared experience pool so that other agents can access and learn from it. Then, centralized-distributed training is performed, where agents perform training locally but share global information to facilitate better collaborative learning. A value function network is utilized to compute the advantage function for each agent. Finally, the policy network is updated for each agent until a satisfactory level of performance is reached.

Initialize θ_i, ϕ_i for the actor−critic network
Set learning rates α^{θ}, α^{ϕ}
For episode (i.e., an operating day) $data = 1$ to E
Initialize both the local observation $s_{i,\theta}$ and global state s_{θ}
For each time step (e.g., 1 hour) $t = 1$ to T
For VPP agents, $i = 1$ to N **do**
Chooses PPO action $a_{i,t}$ according to observation $s_{i,t}$ via the policy $\pi_{\theta i}(a|s)$
Correct action $a_{i,t}$ of all agent values based on the security layer
Observes reward $r_t{}^s$ and the next observation $s_{i,t+1}$
Stores the sample experience into trajectory $\tau_i += [s_{i,t}, a_{i,t}{}^s, r_t{}^s]$
End for
Updates observation $s_{i,t} \rightarrow s_{i,t+1}$ for the VPP agent i
End for
For VPP agents, $i = 1$ to N
Approximates discounted reward-to-go $\hat{r}_t$ r and advantage function $\hat{A}_{i,t}$ utilizing trajectory τ_i
Updates the parameters θ_i, ϕ_i of networks in (49)−(50)
End for
End for

Algorithm 2. Multi-agent GRUPPO for agents.

4 Case studies

In this section, case studies are conducted to show the effectiveness and advantages of the proposed GRUPPO approach for the VPP economic dispatch. The simulation tests are undertaken based on the actual historical data, which are compared with the other two schemes: the PPO scheme and the non-economic dispatch scheme. The detailed parameters of electrical generation, storage, and consumption can be found in Xu et al. (2020), Wang et al. (2023), and Xu et al. (2023).

4.1 Comparison of the convergence and stability performance

In order to compare the stability and convergence, the GRUPPO and PPO algorithms are implemented to optimize the VPP economic dispatch. In order to avoid the randomness of the test results, 10 different random seeds are used to conduct 1,000 rounds. In order to capture the uncertainties of PV power and base load, the Monte Carlo method is implemented to obtain 1,000 scenarios for simulation, where forecasting errors were assumed to follow a normal distribution function. Subsequently, the optimization results of 10 groups of economic dispatch are recorded and depicted in Figure 4. The mean variances of the corresponding 1,000 rounds are also calculated to further explain the differences in stability and convergence performance.

It can be seen in Figure 4 that both methods can achieve almost stable rewards after approximately 30 rounds. However, the reward in the PPO scheme shows a significant increase after 420 rounds in the test, i.e., from approximately −4,600 to −4,900. In contrast, the reward of the proposed GRUPPO still fluctuates up and down near −4,600. Thus, it can be concluded that the GRUPPO approach has better convergence and is more stable to optimize the

VPP economic dispatch. These results also illustrate that the introduction of the GRU network into PPO can fully consider the historical time series information and effectively improve the performance of the PPO algorithm.

4.2 Comparison of the VPP economic dispatch

Based on the actual data, the VPP economic dispatch results are tested using GRUPPO and PPO approaches. The general supply–demand results under three schemes are shown in Figures 5–7, and Figures 8–11 depict the detailed operational results of VPP components. Compared with other two schemes, the battery energy storage under the GRUPPO approach can appropriately store the excess photovoltaic power for later release. It can be observed that the flexible load of the VPP can increase its demand as the PV generation increases. In contrast, in the non-economic dispatch scenario, these loads are evenly distributed over the 24-h period. Although the PPO method can also dispatch all components, the flexible loads do not exhibit higher demand during the high PV generation for 10–15 h. This would increase the pressure on the thermal power units and the power purchase cost.

It can be seen in Figure 10 that the power generation of the thermal power units in the GRUPPO method shows more intense fluctuations compared to that in the other two methods. This indicates that the proposed method can better adjust the thermal power generation to follow the changes in PV power, thereby reducing the VPP operating costs. Moreover, it can be observed from Figure 11 that the HVAC power increases with the increase in PV power. In contrast, in the absence of the economic dispatch, the HVAC power changes with the daytime temperature. Although PPO can also dispatch the HVAC power to follow the power fluctuation, its sensitivity is lower than that of the GRUPPO method.

The overall operating costs of three schemes are 4,322$, 4,431$, and 4,620$, respectively. It is evident that the operating cost of the GRUPPO method is the lowest. Specifically, compared to the PPO method and non-economic dispatch scheme, the proposed GRUPPO method reduces the operating costs by 2.4% and 6.5%, respectively. Overall, these results demonstrate the effectiveness and superiority of the proposed GRUPPO method in reducing the VPP economic dispatch costs.

5 Conclusion

This paper proposed a deep reinforcement learning-based VPP economic dispatch framework. The VPP economic dispatch is captured via a POMDP, which is then solved using a novel GRUPPO approach. The findings of this paper are summarized as follows:

(1) Compared with PPO, the proposed GRUPPO approach can make full use of the time series characteristics, improving its convergence and stability performance.
(2) Based on the POMDP, GRUPPO learns to make decisions based on partial information, which is suitable to handle real-world optimization problems with uncertainty.
(3) Both continuous and discrete actions can be effectively handled using the proposed GRUPPO approach, thereby achieving a maximum cost reduction of 6.5%.
(4) The GRUPPO strategy can outperform other methods in higher economy and scalability, exhibiting huge development and application potentialities in the high-renewable modern power system.

Electrical market development has become an inevitable trend under the background of economic globalization and industrial revolution. Further research would focus on strategic offering of the VPP in the electrical market.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material; further inquiries can be directed to the corresponding author.

Author contributions

ZG: investigation, project administration, writing–original draft, writing–review and editing, and conceptualization. WK: formal analysis, funding acquisition, methodology, writing–original draft, writing–review and editing, and conceptualization. XC: conceptualization, methodology, visualization, writing–original draft, and writing–review and editing. SG: data curation, methodology, writing–original draft, and writing–review and editing. ZL: conceptualization, formal analysis, investigation, writing–original draft, and writing–review and editing. DH: conceptualization, formal analysis, investigation, methodology, writing–original draft, writing–review and editing, and resources. SS: writing–original draft, writing–review and editing, investigation, methodology, software, validation, and visualization. X-CS: conceptualization, investigation, supervision, writing–original draft, writing–review and editing, and validation.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. The authors gratefully acknowledge the support of the Hubei Provincial Natural Science Foundation of China under Grant 2022CFB907.

Conflict of interest

Authors ZG, WK, XC, SG, and ZL were employed by Hubei Branch of State Power Investment Group Co., Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Bui, V., Hussain, A., and Kim, H. M. (2020). Double deep Q -Learning-Based distributed operation of battery energy storage system considering uncertainties. *IEEE Trans. Smart Grid* 11, 457–469. doi:10.1109/TSG.2019.2924025

Canizo, M., Triguero, I., Conde, A., and Onieva, E. (2019). Multi-head CNN–RNN for multi-time series anomaly detection: an industrial case study. *Neurocomputing* 363, 246–260. doi:10.1016/j.neucom.2019.07.034

Chen, G., and Li, J. (2018). A fully distributed ADMM-based dispatch approach for virtual power plant problems. *Appl. Math. Model.* 58, 300–312. doi:10.1016/j.apm.2017.06.010

Du, L., Zhang, L., Tian, X., and Lei, J. (2018). Efficient forecasting scheme and optimal delivery approach of energy for the energy Internet. *IEEE Access* 6, 15026–15038. doi:10.1109/ACCESS.2018.2812211

Etherden, N., Vyatkin, V., and Bollen, M. H. J. (2015). Virtual power plant for grid services using IEC 61850. *IEEE Trans. Ind. Inf.* 12, 437–447. doi:10.1109/TII.2015.2414354

Gough, M., Santos, S. F., Lotfi, M., Javadi, M. S., Osorio, G. J., Ashraf, P., et al. (2022). Operation of a technical virtual power plant considering diverse distributed energy resources. *IEEE Trans. Ind. Appl.* 58, 2547–2558. doi:10.1109/TIA.2022.3143479

Gronauer, S., and Diepold, K. (2023). Multi-agent deep reinforcement learning: a survey. *Artif. Intell. Rev.* 55, 895–943. doi:10.1007/s10462-021-09996-w

Hua, H., Qin, Y., Hao, C., and Cao, J. (2019). Optimal energy management strategies for energy Internet via deep reinforcement learning approach. *Appl. Energy.* 239, 598–609. doi:10.1016/j.apenergy.2019.01.145

HuangYangZhang, S. M. C., Gao, Y., and Yun, J. (2021). A control strategy based on deep reinforcement learning under the combined wind-solar storage system. *IEEE Trans. Ind. Appl.* 57, 6547–6558. doi:10.1109/TIA.2021.3105497

Koraki, D., and Strunz, K. (2017). Wind and solar power integration in electricity markets and distribution networks through service-centric virtual power plants. *IEEE Trans. Power Syst.* 33, 473–485. doi:10.1109/TPWRS.2017.2710481

Książek, W., Abdar, M., Acharya, U. R., and Pławiak, P. (2019). A novel machine learning approach for early detection of hepatocellular carcinoma patients. *Cogn. Syst. Res.* 54, 116–127. doi:10.1016/j.cogsys.2018.12.001

Lin, L., Guan, X., Peng, Y., Wang, N., Maharjan, S., and Ohtsuki, T. (2020). Deep reinforcement learning for economic dispatch of virtual power plant in internet of energy. *IEEE Internet Things J.* 7, 6288–6301. doi:10.1109/JIOT.2020.2966232

Liu, L., Xu, D., and Lam, C. S. (2023). Two-layer management of HVAC-based multi-energy buildings under proactive demand response of fast/slow-charging EVs. *Energy Convers. Manag.* 289, 117208. doi:10.1016/j.enconman.2023.117208

Liu, W., Zhang, D., Wang, X., and Hou, J. (2023). A decision making strategy for generating unit tripping under emergency circumstances based on deep reinforcement learning. *Proc. CSEE* 38, 109–119. doi:10.13334/j.0258-8013.pcsee.171747

Liu, Y. Y., Li, M., Lian, H., Tang, X., Liu, C., and Jiang, C. (2018). Optimal dispatch of virtual power plant using interval and deterministic combined optimization. *Int. J. Electr. Power Energy Syst.* 102, 235–244. doi:10.1016/j.ijepes.2018.04.011

Lu, R., and Hong, S. H. (2019). Incentive-based demand response for smart grid with reinforcement learning and deep neural network. *Appl. Energy.* 236, 937–949. doi:10.1016/j.apenergy.2018.12.061

Ma, Y., Hu, Z., and Song, Y. (2023). A Reinforcement learning based coordinated but differentiated load frequency control method with heterogeneous frequency regulation resources. *IEEE Trans. Power Syst.* 39, 2239–2250. doi:10.1109/TPWRS.2023.3262543

Mashhour, E., and Moghaddas-Tafreshi, S. M. (2010). Bidding strategy of virtual power plant for participating in energy and spinning reserve markets—Part I: problem formulation. *IEEE Trans. Power Syst.* 26 (2), 949–956. doi:10.1109/TPWRS.2010.2070884

Naveen, R., Revankar, P. P., and Rajanna, S. (2020). Integration of renewable energy systems for optimal energy needs-a review. *Int. J. Energy Res.* 10, 727–742. doi:10.20508/ijrer.v10i2.10571.g7944

Thanh, P., Cho, M., Chang, C. L., and Chen, M. J. (2022). Short-term three-phase load prediction with advanced metering infrastructure data in smart solar microgrid based convolution neural network bidirectional gated recurrent unit. *IEEE Access* 10, 68686–68699. doi:10.1109/ACCESS.2022.3185747

Wang, Y., Qiu, D., Sun, X., Bie, Z., and Strbac, G. (2023). Coordinating multi-energy microgrids for integrated energy system resilience: a multi-task learning approach. *IEEE Trans. Sustain. Energy*, 1–18. doi:10.1109/TSTE.2023.3317133

Xu, D., Wu, Q., Zhou, B., Bai, L., and Huang, S. (2019). Distributed multi-energy operation of coupled electricity, heating, and natural gas networks. *IEEE Trans. Sustain. Energy* 11, 2457–2469. doi:10.1109/TSTE.2019.2961432

Xu, D., Zhou, B., Wu, Q., Chung, C. Y., Huang, S., Chen, S., et al. (2020). Integrated modelling and enhanced utilization of power-to-ammonia for high renewable penetrated multi-energy systems. *IEEE Trans. Power Syst.* 35, 4769–4780. doi:10.1109/TPWRS.2020.2989533

Xu, D., Zhou, B., Liu, N., Wu, Q., Voropai, N., Li, C., et al. (2021). Peer-to-peer multienergy and communication resource trading for interconnected microgrids. *IEEE Trans. Ind. Inf.* 17, 2522–2533. doi:10.1109/TII.2020.3000906

Xu, D., Zhong, F., Bai, Z., Wu, Z., Yang, X., and Gao, M. (2023). Real-time multi-energy demand response for high-renewable buildings. *Energy Build.* 281, 112764. doi:10.1016/j.enbuild.2022.112764

Yang, S., Tan, S., and Xu, J. X. (2013). Consensus based approach for economic dispatch problem in a smart grid. *IEEE Trans. Power Syst.* 28, 4416–4426. doi:10.1109/TPWRS.2013.2271640

Zhang, Y., Ai, X., Wen, J., Fang, J., and He, H. (2018). Data-adaptive robust optimization method for the economic dispatch of active distribution networks. *IEEE Trans. Smart Grid* 10 (4), 3791–3800. doi:10.1109/TSG.2018.2834952

Zhao, L., Yang, T., Li, W., and Zomaya, A. Y. (2022). Deep reinforcement learning-based joint load scheduling for household multi-energy system. *Appl. Energy.* 324, 119346. doi:10.1016/j.apenergy.2022.119346