Electrical Power and Energy Systems 158 (2024) 109980

Contents lists available at ScienceDirect

International Journal of Electrical Power and Energy Systems

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

Federated reinforcement learning for Short-Time scale operation of Wind-Solar-Thermal power network with nonconvex models

Yao Zou [a], Qianggang Wang [a], Qinqin Xia [a], Yuan Chi [a,*], Chao Lei [a,b], Niancheng Zhou [a]

[a] State Key Laboratory of Power Transmission Equipment Technology, School of Electrical Engineering, Chongqing University, China
[b] Centre for Advances in Reliability and Safety and Department of Electrical Engineering, Hong Kong Polytechnic University, Kowloon, Hong Kong, China

ARTICLE INFO

Keywords:
Federated reinforcement learning
Renewable energy
Thermal power unit
Power sources scheduling

ABSTRACT

To schedule power sources operated by different entities in a short-time scale considering nonconvex generation cost and deep peak regulation (DPR) service constraints, this paper proposes an FRL-based multiple power sources coordination framework in wind-solar-thermal power network. In the studied power transmission network (TN), renewable energy sources and thermal power units connected to the same bus are aggregated as a wind-solar-thermal virtual power plant (WSTVPP). The transmission system operator (TSO) sends dispatch instructions to each WSTVPP by optimal power flow program, and allocates the cost of DPR service in TN. Based on the dispatch instruction, the internal power sources of each WSTVPP are scheduled by its local center control agent to achieve local economic operation while maximizing the overall DPR service revenue for the WSTVPP from the auxiliary service market. The multiple WSTVPPs operation is modeled as a partially observable Markov decision process, and solved by a designed FRL algorithm. The FRL algorithm employs a global neural network (NN) model for coordination, heterogeneous local NN models and data to efficiently train each WSTVPP control agent with individual objectives for handling multiple power sources scheduling in TN while preserving local privacy. Numerical studies validate the effectiveness of the proposed framework for handling the short-time scale power sources operation with nonconvex constraints.

1. Introduction

INTEGRATION of high-penetration renewable energy sources (RESs) into the power grid have significantly increased the demand for flexible power sources [1,2]. The stochastic fluctuations of PV and WT generations give rise to the unpredictability and instability of power systems [3]. By aggregating or deploying the RESs and flexible power sources at close distance in the form of a virtual power plant (VPP) [4], the RES can directly take advantage of flexible power sources to handle their fluctuation locally and enhance efficiency.

On the generation side, thermal power units (TPUs) play a central role as the main flexible and extensively utilized power sources within TN in regions with limited water resources [5]. However, the flexibility of TPUs is restricted by their inherent physical characteristics. To address this issue, flexibility retrofit of TPUs is a straightforward and realistic method. For example, by setting up a dust bunker between the coal mill and the burner to store pulverized coal, the minimum output of TPU can be lowered to accommodate more RES [5,6], and gain more benefits through real-time DPR auxiliary services [7]. On the other hand, the complexity of TN with multiple VPPs is increased due to the presence of more components, higher communication overheads, simultaneous local behaviors, and nonconvex operation characteristics on a short-time scale. For example, research in [8] demonstrates that TPU has different maximum ramp rates under different output statuses, which may lead to nonconvex ramping constraints on the 15-minute time scale. Furthermore, as the integration of RES increases, optimization of VPP operation considering real-time DPR auxiliary services on such a short time scale is also essential for a flexible power source to balance the stochastic load demand and RES generation accurately and sufficiently [9]. However, relevant research has not thoroughly considered the aforementioned nonconvex characteristics [10–12]. Therefore, the design of an innovative scheduling architecture, which can effectively aggregate RESs with retrofitted TPUs in VPP and coordinate VPPs in TN, brings about a new challenge.

The TN and its internal VPPs are usually operated by different operators with distinguished objectives [13], coordination and communication are imperative among different power source operators and transmission system operator (TSO). Privacy preservation is also a challenge that need to be addressed in TN operation. Both TSO and

* Corresponding author.
E-mail address: chiyuanee@cqu.edu.cn (Y. Chi).

https://doi.org/10.1016/j.ijepes.2024.109980
Received 27 November 2023; Received in revised form 9 March 2024; Accepted 7 April 2024
Available online 15 April 2024

Nomenclature

Indices

$g \in G$	Index and set of generation groups in transmission network (TN).
$i \in \mathscr{I}$	Index, set of wind-solar-thermal virtual power plants (WSTVPPs) and their control agent.
$m \in M$	Index and set of emission type
$t \in T$	Index and set of time steps.
$u \in U_i$	Index and set of thermal power units (TPUs).
Δt	Time interval. (1/4 h, i.e. 15 min)

Superscript

TP	Thermal Power generation.
N/D	Normal/Deep peak regulation states.
O/E	Operation/Emission.
B/F	Basic/Flexibility active output
WT	Wind Turbine generation.
PV	PhotoVoltaics generation.
c	Corrected output.
P	Penalty factor.
R	Revenue.
S	Shared cost.
TN	Transmission Network.

Parameters (a) Thermal power

$A^{TP,S}_{1/2}$	Load rate standard for deep peak regulation (DPR) service shared cost of TPU.
$a^{TP,D}_{i,u}$, $b^{TP,D}_{i,u}$	Fitting coefficients of TPU DPR loss. (ton/MW, ton)
$a^{TP,E}_{i,u}$, $b^{TP,E}_{i,u}$	Fitting coefficients of TPU pollution. (m3/MW, m3)
$a^{TP,O}_{i,u}$, $b^{TP,O}_{i,u}$, $c^{TP,O}_{i,u}$	Fitting coefficients of TPU operation cost. (ton/MW2, ton/MW, ton)
$C^{TP,R}_{0/1/2}$	Unit generation revenue of TPUs in different compensation standard. (CNY/MW)
C^{coal}	Unit cost of coal. (CNY/ton)
$C^{TP,E}_{m}$	Unit generation cost of emission. (CNY/kg)
$k^{TP,D}_{t}$	Season coefficient for DPR revenue of TPUs.
$\overline{P}^{TP}_{i,u}$	Maximum active power of TPU. (MW)
$\underline{P}^{TP,N/TP,D}_{i,u}$	Minimum active power of TPU at NPR (normal peak regulation), DPR state. (MW)
$R^{TP,N}_{i,u}$, $R^{TP,D}_{i,u}$	Ramping rate of TPU in NPR, DPR states. (MW/h)
$S^{TPU}_{i,u}$	Capacity of TPU. (MVA)
$z^{TP,S}_{1/2/3}$	Corrected factors of DPR shared cost for TPU.
$\mu^{TP,D}_{1/2}$	Load rate standard for DPR compensation revenue of TPU.
$\tau^{TP}_{i,u}$	Real time flexibility regulation rate of TPU.
ρ_m	Emission factors of pollutants. (kg/m3)
σ_m	Equivalent value of pollutants.

(b) Wind and Photovoltaics

$C^{PV/WT,R}$	Unit generation cost of photovoltaic (PV), wind turbine (WT) generation. (CNY/MW)
$C^{PV/WT,P}$	Penalty factor of PV, WT curtailment.
$S^{PV/WT}_{i}$	Capacity of PV/WT generation. (MVA)
$z^{PV,S/WT,S}_{t}$	Corrected factors of DPR shared cost for PV, WT.
τ^{PV}_{i}, τ^{WT}_{i}	Minimum PV, WT utilization rate.

(c) Others

R_g	Overall ramping rate of integrated grid-connected power source g. (MW/h)
S_g	Capacity of integrated grid-connected power source g. (MVA)
$w^{UN,P}$	Penalty factor of unbalanced power. (CNY/MW)

Variables (a) Thermal power

$\overline{a}^{TP}_{i,u,t}$, $\underline{a}^{TP}_{i,u,t}$	Upper, lower bound of active power feasible range for TPU generation. (MW)
$f^{TP}_{i,u,t}$	Objective of TPU.
$P^{TP}_{i,u,t}$	Actual TPU active output. (MW)
$P^{TP,B}_{i,u,t}$	Basic active output of TPU. (MW)
$P^{TP,F}_{i,u,t}$	Real-time active flexibility of TPU. (MW)
$P^{TP,c}_{i,u,t}$	Corrected TPU output for DPR shared cost. (MW)
$\mu^{TP}_{i,u,t}$	Load rate of TPU.

(b) Wind and Photovoltaics

$\overline{a}^{PV/WT}_{i,t}$, $\underline{a}^{PV/WT}_{i,t}$	Upper, lower bound of active power feasible range for PV/WT generation. (MW)
$f^{WT/PV}_{i,t}$	Objective of WT/PV.
$P^{PV*}_{i,t}$, $P^{PV}_{i,t}$	Predict/actual PV active power. (MW)
$P^{WT*}_{i,t}$, $P^{WT}_{i,t}$	Predict/actual WT active power. (MW)
$P^{PV,c}_{i,t}$, $P^{WT,c}_{i,t}$	Corrected PV, WT output for DPR shared cost. (MW)

(c) Others

$C^{TN,D}_{t}$	Total DPR service revenue in TN. (CNY)
f^{OPF}_{t}	Objective of optimal power flow (OPF).
$f^{VPP}_{i,t}$	Objective of WSTVPP.
$\overline{P}_{g,t}$, $\underline{P}_{g,t}$	Upper and lower bound of active power feasible range for g. (MW)
$\overline{Q}_{g,t}$, $\underline{Q}_{g,t}$	Upper and lower bound of reactive power feasible range for g. (MVar)
$P_{g,t}$	Total active power generation of a bus. (MW)
$Q_{g,t}$	Total reactive power generation of a bus. (MVar)
$P^{c}_{g,t}$	Total corrected output of DPR shared cost. (MW)
$P^{L}_{i,t}$	Active dispatch instruction. (MW)
$P^{NET}_{i,t}$	Net active load. (MW)
$P^{UN}_{i,t}$	Unbalanced active power between total generation and dispatch instruction. (MW)

power source operators may be reluctant to share sensitive privacy information since the detailed power output information may reveal the behavior of the power source operators, potentially leading to privacy disclosures and unfair competition [14]. For example, peak regulation strategy and technical characteristics of power sources could be inferred from the output data of each power source in a VPP, enabling intentional operators to devise targeted strategies for gaining undue advantages [15].

Traditional TPUs or multiple VPPs scheduling in power system operation is typically formulated as a model-based optimization problem to maximize the total generation revenue of the system. When dealing with a large-scale power grid with multiple power sources, the prevalent centralized optimization methods [16,17] adopt a center coordinator for modeling, data processing, and calculation of power sources scheduling. These methods suffer from heightened communication and computational demands, as well as concerns about privacy. To address these challenges, distributed or decentralized optimization methods [18–21] have been developed to break down the global

optimization problem into several subproblems that can be solved locally. However, these methods cannot handle the nonlinear or nonconvex problem effectively. Apart from the aforementioned short-time scale ramping constraints of TPUs, generation cost functions and revenue sharing mechanism of DPR service among power sources also have nonlinear or nonconvex characteristics [7,22], which will further pose challenges to the model-based optimization methods.

Reinforcement learning (RL) provides a feasible model-free solution for complex decision-making problems [23] and has been successfully applied in VPP operation problems and nonconvex power sources scheduling problems, such as managing distributed energy resources within VPP for regulation service [24] and the efficient multi-timescale bidding for hybrid power plants [25]. These methods are based on centralized training with all the data collected and processed by a center coordinator, which will inevitably lead to privacy concerns, heavy communication and computation burden. To address the above issues, distributed RL methods with local training and data exchange between neighbors, which reduce communication and computation burden, are developed [26,27]. However, they are still limited due to the complex communication mechanism, insufficient scalability, and direct raw data sharing with neighbors.

To overcome the shortages of the RL methods above, federated reinforcement learning (FRL) has been developed recently to enable distributed local training with privacy preservation. FRL can achieve coordination and efficient training among agents through a global model which is updated by model parameter exchange instead of raw data [28]. FRL has been successfully applied in economic operation of power systems [29,30]. [31] proposes a robust FRL approach to schedule VPPs with electric vehicles and RES, it enables multiple users for training a shared policy model. In [32], an FRL-based algorithm is proposed for decentralized voltage control of multiple VPPs in distribution network. However, the above FRL-based studies mainly adopt the shared neural network (NN) models [29–31] or an incomplete privatization setting [32]. Under the complex multiple VPPs operation environment characterized by nonlinear and nonconvex constraints, each VPP has individual local status. The FRL methods with shared NN models prove to be challenging in achieving good performance with model privacy preservation.

To sum up, several research gaps need to be addressed in the existing studies for multiple VPPs operation in TN: (i) Model-based optimization methods [18–21] require the detailed parameterized model, making them unsuitable for solving the short-time scale large nonconvex problem with high computation efficiency and good convergence. (ii) Centralized-based RL frameworks [24,25] have a heavy communication burden. Besides, they fail to adequately protect the privacy of models and data for power sources operated by different entities, posing a risk of potential exposure of sensitive generation information, control algorithms, and strategies. (iii) Distributed RL methods [26,27] necessitate specifically designed communication methods, resulting in compromises in scalability and efficiency for both maintenance and training. (iv) Most FRL methods are based on shared NN models or incomplete privatization settings [29–32], which struggle to effectively coordinate the complex nonconvex operation objective of individual VPP and balance the power supply and demand with good performance. (v) Previous studies for the operation of multiple VPPs mainly focus on the distribution side, little or no research has been implemented on the generation side considering DPR auxiliary service.

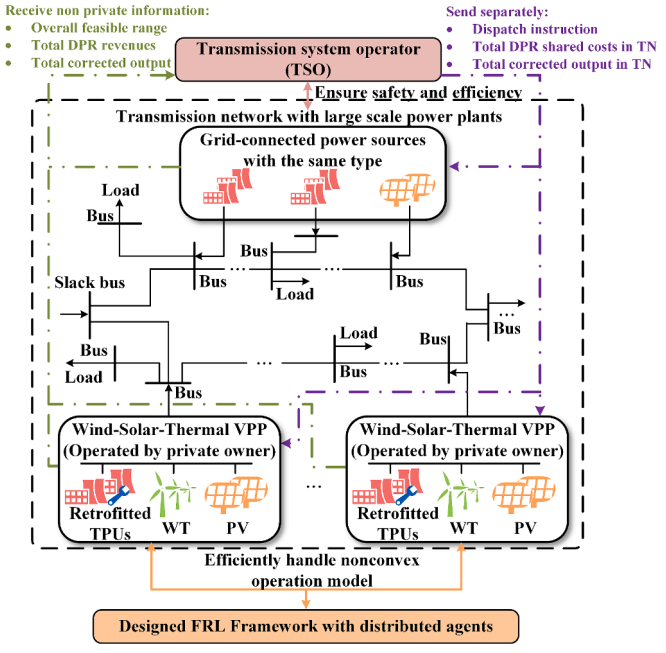


Fig. 1. Overall framework of the studied system.

To fill the aforementioned research gaps, this paper proposes a hybrid approach that combines data-driven FRL with the model-based method for the coordinated operation of multiple power sources in a wind-solar-thermal TN. The proposed framework aggregated the grid-connected RES and TPUs at the same bus in TN as several WSTVPPs. The TN and WSTVPPs are operated separately with limited necessary information exchange to achieve coordination while preserving privacy. Power sources in each WSTVPP are scheduled by individual control agents for local economic operation considering TN DPR auxiliary service. The designed FRL algorithm with both heterogeneous local NN models for each agent and the global NN model for coordination is used to efficiently train the WSTVPP control agents. During training, the proposed FRL algorithm can enhance the training performance, and preserve the local models and data privacy by exchanging only the parameters of the global NN model. The contributions of this paper are summarized as follows:

1) A hybrid approach that integrates the data-driven FRL into the model-based method is proposed to address the scheduling of diverse power sources within a TN considering the participation of DPR auxiliary services. The proposed approach breaks down the complex scheduling challenge into distinct TN OPF and WSTVPP operation models while ensuring minimal yet effective information exchange (limited to dispatch and DPR service cost data).

2) The WSTVPPs operation with a strategic information exchange mechanism for privacy and coordination is reformulated as a decentralized partially observable Markov decision process (Dec-POMDP) considering the power sources characteristics. This transformation can enable the scheduling of internal power sources by WSTVPP control agents for effectively following dispatch instructions, accommodating RES generation, and achieving economic operation.

3) A customized FRL algorithm incorporating a global coordinator and NN model alongside heterogeneous local NN models for distributed agents is proposed. The algorithm can efficiently address the nonconvex challenges of multiple WSTVPPs coordinated operation in TN, while safeguarding local NNs' privacy through parameter exchanges of the global NN model, handling the individual agent objective to achieve local economic operation, obtain DPR service revenue from TN, and track load demand.

2. Multiple Power Sources Operation Modeling of Wind-Solar-Thermal Power Transmission Network

In this section, the overall TN and multiple power sources operation framework is discussed first. Then the detailed operation model for TN OPF and WSTVPP operation considering DPR auxiliary service is presented.

2.1. Problem formulation

In this paper, a TN incorporating several megawatt-level doubly fed induction generator (DFIG) based wind farms, solar farms, and thermal power plants with a capacity of several hundred megawatts is considered. These large-scale power source plants are directly connected to the main bus in TN. Solar farms and DFIG-based wind farms can leverage their inverters to provide adequate reactive power support to TN [33]. Thermal power plants can also provide certain reactive power support to TN [34]. The TSO is responsible for the operation of TN to balance power supply and demand. Several power source groups connected to certain buses are operated by different operators. To dispatch each power source in TN, the TSO needs to access necessary detailed models and data of power sources owned by various operators, leading to privacy concerns. Furthermore, DPR auxiliary service in TN involves complex revenue and sharing mechanisms with nonlinear characteristics, which presents challenges when employing traditional model-based optimization approaches for effective power dispatch. To address these issues, a hybrid approach that combines data-driven FRL with the model-based method for multiple power sources coordinated operation is proposed. The proposed framework enables both privacy preservation and efficient solutions for the separate operation of TN and power sources.

As depicted in Fig. 1, in the TN, the power sources connected to specific buses are regarded as several generation groups. The TSO assigns short-time scale (15 min) dispatch instructions to each generation group based on the OPF program to regulate bus voltage within the allowable range, balance the power supply and demand, and reduce the additional power generation in TN. The generation groups need to follow the dispatch instructions, schedule their internal power sources to supply active and reactive power. The generation groups only need to provide their overall feasible active and reactive output range to TSO for future OPF calculation and dispatch instruction. Additionally, the TSO and the generation groups also exchange information for revenue and shared cost calculation associated with DPR auxiliary service. Specifically, generation groups provide their aggregated DPR compensation revenue and corrected generation output to TSO, the total DPR compensation revenue is subsequently shared among the power sources in TN. The TSO leverages the above information from generation groups to calculate the specific shared costs of each generation group and send them back to respective generation groups for local economic calculation.

In the TN, certain generation groups comprise power sources of the same type with similar economic and technical characteristics. These groups can easily coordinate the output of their internal power sources in accordance with the active and reactive power dispatch instructions, which can be regarded as an integrated grid-connected power source (represented as $g \notin I$). On the other hand, there are generation groups involve RES and nearby high performance retrofitted TPUs provide flexible resources with varying characteristics. These groups face the challenge of considering the intricate complex economic and technical characteristics of different types of power sources, along with the need to incorporate DPR auxiliary services into their internal power source scheduling. In this paper, the generation group with different types of power sources (i.e., WT, PV, and TPUs) is considered as a WSTVPP (represented as $(g = i) \in I$), which internal power sources are scheduled by its local center controller based on the active and reactive power dispatch instructions from TSO. To accommodate more RES by flexible power sources and collaborate to gain more revenue, it is reasonable for internal power sources in a WSTVPP to share their generation information to the local center controller. The WSTVPP controller acts as a control agent. The control agent of WSTVPP can directly transmit power generation instructions to the automatic generation control system of the power sources within the VPP, thereby controlling the output of the power sources. The control agents of WSTVPP in TN are efficiently trained by a FRL algorithm to achieve global interaction with TN and local economic operation with privacy preservation.

The proposed framework aims to address several critical challenges in multiple power sources operation in TN. Specifically: (i) The large-scale nonlinear and nonconvex short-time scale operation problem of TN considering DPR auxiliary service is decomposed in TN level and power source group level. (ii) The TSO does not require the detailed model of each power source for OPF calculation, which enables efficient separate operation and privacy preservation. (iii) The WSTVPP operation considering both global DPR auxiliary service and local economic operation can be iteratively trained by the designed FRL algorithm to avoid solving complex and time-consuming nonlinear optimization problem.

2.2. Optimal power flow in transmission network

At the TN level, the objective of TN OPF in this paper is to minimize the overall generation feed-in at each time step:

$$\min f_t^{\mathrm{OPF}} = \sum_{g \in G} P_{g,t}. \tag{1}$$

The constraints include commonly used polar active/reactive power balance of buses, active/reactive power injection equations of buses between generation and load, voltage limits of buses, and branch transmission limits, of which details can be found in studies of optimal power flow analysis, such as [35]. Other constraints are listed as follows:

$$\underline{P}_{g,t} \le P_{g,t} \le \overline{P}_{g,t}, \forall g \in G, \tag{2}$$

$$\underline{Q}_{g,t} \le Q_{g,t} \le \overline{Q}_{g,t}, \forall g \in G, \tag{3}$$

$$\begin{cases} \underline{P}_{g,t} = P_{g,t-1} - R_g \Delta t \\ \overline{P}_{g,t} = P_{g,t-1} + R_g \Delta t \end{cases}, \forall g \notin I, \tag{4}$$

$$\begin{cases} \underline{P}_{g,t} = \sum_{u \in U_i} \underline{a}_{i,u,t}^{\mathrm{TP}} + P_{i,t}^{\mathrm{WT}*} + P_{i,t}^{\mathrm{PV}*} \\ \overline{P}_{g,t} = \sum_{u \in U_i} \overline{a}_{i,u,t}^{\mathrm{TP}} + P_{i,t}^{\mathrm{WT}*} + P_{i,t}^{\mathrm{PV}*} \end{cases}, \forall (g = i) \in I, \tag{5}$$

$$\overline{Q}_{g,t}, \underline{Q}_{g,t} = \pm \sqrt{S_g^2 - \overline{P}_{g,t}^2}, \forall g \notin I, \tag{6}$$

$$\begin{aligned} \overline{Q}_{g,t}, \underline{Q}_{g,t} = \pm \Bigg[& \sum_{u \in U_i} \sqrt{\left(S_{i,u}^{\mathrm{TP}}\right)^2 - \left(\overline{a}_{i,u,t}^{\mathrm{TP}}\right)^2} + \sqrt{\left(S_i^{\mathrm{PV}}\right)^2 - \left(P_{i,t}^{\mathrm{PV}*}\right)^2} \\ & + \sqrt{\left(S_i^{\mathrm{WT}}\right)^2 - \left(P_{i,t}^{\mathrm{WT}*}\right)^2} \Bigg], \forall (g = i) \in I, \end{aligned} \tag{7}$$

where (2), (3) are the active, reactive power feasible range of generation units or WSTVPP, where $\underline{P}_{g,t}$, $\overline{P}_{g,t}$, $\underline{Q}_{g,t}$, $\overline{Q}_{g,t}$ are calculated by corresponding grid-connected power source or WSTVPP and informed to TSO in this paper. The active power feasible ranges of grid-connected power source and WSTVPP are constrained by (4) and (5). (5) considers the real-time output of RES, ramping constraints, and output of TPUs at the last time step in WSTVPP, where $\underline{a}_{i,u,t}^{\mathrm{TP}}$ and $\overline{a}_{i,u,t}^{\mathrm{TP}}$ are designed in the FRL framework by eqs. (27) and (28), the details will be discussed in Sections II-C and III. The reactive power feasible ranges are constrained by power source capacities and maximum active feasible output of power sources. For grid-connected power source and WSTVPP, the reactive power feasible range can be calculated by (6) and (7) respectively. (7) is designed to ensure that the active output of power sources in WSTVPP will not affect its overall reactive power capability.

2.3. Operation model of virtual power plant

The objective for the operation of WSTVPP i is to maximize the overall revenue including its DPR auxiliary service revenue and shared

cost generated from global interaction with TN, local economic operation, and reduce the unbalanced active power between generation and dispatching instruction at t:

$$\max f_{i,t}^{\mathrm{VPP}} = \Delta t\left(\sum_{u\in U_i} f_{i,u,t}^{\mathrm{TP}} + f_{i,t}^{\mathrm{PV}} + f_{i,t}^{\mathrm{WT}} - w^{\mathrm{UN,P}}\left|P_{i,t}^{\mathrm{UN}}\right|\right). \tag{8}$$

The overall generation revenue of TPUs $f_{i,u,t}^{\mathrm{TPU}}$ is

$$f_{i,u,t}^{\mathrm{TP}} = f_{i,u,t}^{\mathrm{TP,R,N}} + f_{i,u,t}^{\mathrm{TP,R,D}} - f_{i,u,t}^{\mathrm{TP,O}} - f_{i,u,t}^{\mathrm{TP,E}} - f_{i,u,t}^{\mathrm{TP,S}}, \tag{9}$$

which includes coal consumption/operation cost $f_{i,u,t}^{\mathrm{TP,O}}$, environment cost $f_{i,u,t}^{\mathrm{TP,E}}$, basic revenue $f_{i,u,t}^{\mathrm{TP,R,N}}$ DPR auxiliary service compensation revenue $f_{i,u,t}^{\mathrm{TP,R,D}}$ and shared cost $f_{i,u,t}^{\mathrm{TP,S}}$,. Where the $f_{i,u,t}^{\mathrm{TP,R,D}}$ can be obtained when the average load rate of TPUs is below the given compensation standard at t [7]. $f_{i,u,t}^{\mathrm{TP,R,N}}, f_{i,u,t}^{\mathrm{TP,R,D}}$ are calculated by

$$f_{i,u,t}^{\mathrm{TP,R,N}} = C_0^{\mathrm{TP,R}} P_{i,u,t}^{\mathrm{TP}}, \tag{10}$$

$$f_{i,u,t}^{\mathrm{TP,R,D}} = \begin{cases} 0, & \mu_1^{\mathrm{TP,D}} \le \mu_{i,u,t}^{\mathrm{TP}} \le 1 \\ k_t^{\mathrm{TP,D}} C_1^{\mathrm{TP,R}}\left(\mu_1^{\mathrm{TP,D}}\overline{P}_{i,u}^{\mathrm{TP}} - P_{i,u,t}^{\mathrm{TP}}\right), & \mu_2^{\mathrm{TP,D}} \le \mu_{i,u,t}^{\mathrm{TP}} < \mu_1^{\mathrm{TP,D}} \\ k_t^{\mathrm{TP,D}} C_2^{\mathrm{TP,R}}\left(\mu_1^{\mathrm{TP,D}}\overline{P}_{i,u}^{\mathrm{TP}} - P_{i,u,t}^{\mathrm{TP}}\right), & 0 \le \mu_{i,u,t}^{\mathrm{TP}} \le \mu_2^{\mathrm{TP,D}} \end{cases}, \mu_{i,u,t}^{\mathrm{TP}}$$

$$= \frac{P_{i,u,t}^{\mathrm{TP}}}{\overline{P}_{i,u}^{\mathrm{TP}}}, \tag{11}$$

where the first line in eq. (11) represents DPR service revenue only can be obtained when the output rate of a TPU $\mu_{i,u,t}^{\mathrm{TP}}$ is lower than $\mu_1^{\mathrm{TP,D}}$. The second and third line in eq. (11) represent the current active output of TPU can obtain first or second-tier DPR service revenue with a unit price of $C_1^{\mathrm{TP,R}}$ or $C_1^{\mathrm{TP,R}}$. The revenue is calculated based on the difference between the active output corresponding to the compensation standard $\mu_1^{\mathrm{TP,D}}\overline{P}_{i,u}^{\mathrm{TP}}$ and the current active output $P_{i,u,t}^{\mathrm{TP}}$. Besides, the DPR service revenue undergoes seasonal adjustments, incorporating a correction coefficient represented by $k_t^{\mathrm{TP,D}}$.

$f_{i,u,t}^{\mathrm{TP,O}}$ is associated with the NPR, DPR states, which comprises coal consumption cost function and an additional cost function required to sustain TPU operation under low output conditions when the TPU is in the DPR state. It can be calculated by [10]

$$f_{i,u,t}^{\mathrm{TP,O}} = \begin{cases} (1)P_{i,u,t}^{\mathrm{TP}} \in \left[\underline{P}_{i,u}^{\mathrm{TP,N}}, \overline{P}_{i,u}^{\mathrm{TP}}\right] \\ C^{\mathrm{coal}}\left[a_{i,u}^{\mathrm{TP,O}}\left(P_{i,u,t}^{\mathrm{TP}}\right)^2 + b_{i,u}^{\mathrm{TP,O}} P_{i,u,t}^{\mathrm{TPU}} + c_{i,u}^{\mathrm{TP,O}}\right] \\ (2)P_{i,u,t}^{\mathrm{TP}} \in \left[\underline{P}_{i,u}^{\mathrm{TP,D}}, \underline{P}_{i,u}^{\mathrm{TP,N}}\right] \\ C^{\mathrm{coal}}\left[a_{i,u}^{\mathrm{TP,O}}\left(P_{i,u,t}^{\mathrm{TP}}\right)^2 + b_{i,u}^{\mathrm{TP,O}} P_{i,u,t}^{\mathrm{TP}} + c_{i,u}^{\mathrm{TP,O}}\right] + a_{i,u}^{\mathrm{TP,D}} P_{i,u,t}^{\mathrm{TP}} + a_{i,u}^{\mathrm{TP,D}} \end{cases} \tag{12}$$

$f_{i,u,t}^{\mathrm{TP,E}}$ can be calculated by

$$f_{i,u,t}^{\mathrm{TP,E}} = \sum_{m=1}^{M} C_m^{\mathrm{TP,E}} \frac{\rho_m}{\sigma_m}\left(a_{i,u}^{\mathrm{TP,E}} \bullet P_{i,u,t}^{\mathrm{TP}} + b_{i,u}^{\mathrm{TP,E}}\right). \tag{13}$$

The compensation revenue of DPR auxiliary service in TN is shared by WT, PV, and TPUs which load rates are higher than the compensation standard in the system. For TPU, the higher its current output, the more cost need to be shared. The shared cost $f_{i,u,t}^{\mathrm{TP,S}}$ is calculated by [7]

$$f_{i,u,t}^{\mathrm{TP,S}} = \begin{cases} C_t^{\mathrm{TN,D}} \bullet P_{i,u,t}^{\mathrm{TP,c}} \Big/ \sum_{g\in G} P_{g,t}^{\mathrm{c}}, & \mu_{i,u,t}^{\mathrm{TP}} \ge \mu_1 \\ 0, & \mu_{i,u,t}^{\mathrm{TP}} < \mu_1 \end{cases}, \tag{14}$$

$$P_{i,u,t}^{\mathrm{TP,c}} = \begin{cases} (1)P_{i,u,t}^{\mathrm{TP}} \in \left[\underline{P}_{i,u}^{\mathrm{TP,R}}, A_1^{\mathrm{TP,S}}\overline{P}_{i,u}^{\mathrm{TP}}\right] \\ z_1^{\mathrm{TP,S}} P_{i,u,t}^{\mathrm{TP}} \\ (2)P_{i,u,t}^{\mathrm{TP}} \in \left[A_1^{\mathrm{TP,S}}\overline{P}_{i,u}^{\mathrm{TP}}, A_2^{\mathrm{TP,S}}\overline{P}_{i,u}^{\mathrm{TP}}\right] \\ z_1^{\mathrm{TP,S}} A_1^{\mathrm{TP,S}}\overline{P}_{i,u}^{\mathrm{TP}} + z_2^{\mathrm{TP,S}}\left(P_{i,u,t}^{\mathrm{TP}} - A_1^{\mathrm{TP,S}}\overline{P}_{i,u}^{\mathrm{TP}}\right) \\ (3)P_{i,u,t}^{\mathrm{TP}} \in \left[A_2^{\mathrm{TP,S}}\overline{P}_{i,u}^{\mathrm{TP}}, \overline{P}_{i,u}^{\mathrm{TP}}\right] \\ z_1^{\mathrm{TP,S}} A_1^{\mathrm{TP,S}}\overline{P}_{i,u}^{\mathrm{TPU}} + z_2^{\mathrm{TP,S}}\left(A_2^{\mathrm{TP,S}} - A_1^{\mathrm{TP,S}}\right)\overline{P}_{i,u}^{\mathrm{TP}} + z_3^{\mathrm{TP,S}}\left(P_{i,u,t}^{\mathrm{TP}} - A_2^{\mathrm{TP,S}}\overline{P}_{i,u}^{\mathrm{TP}}\right) \end{cases}, \tag{15}$$

$$P_{i,t}^{\mathrm{c}} = \sum_{u\in U_i} P_{i,u,t}^{\mathrm{TP,c}} + P_{i,t}^{\mathrm{WT,c}} + P_{i,t}^{\mathrm{PV,c}}, \tag{16}$$

$$C_t^{\mathrm{TN,D}} = \sum_{i\in I}\sum_{u\in U_i} f_{i,u,t}^{\mathrm{TP,D,R}}. \tag{17}$$

The overall generation revenue of PV $f_{i,t}^{\mathrm{PV}}$ and WT $f_{i,t}^{\mathrm{WT}}$ are

$$f_{i,t}^{\mathrm{PV}} + f_{i,t}^{\mathrm{WT}} = C^{\mathrm{PV,R}} P_{i,t}^{\mathrm{PV}} + C^{\mathrm{WT,R}} P_{i,t}^{\mathrm{WT}} - f_{i,t}^{\mathrm{WT,S}} - f_{i,t}^{\mathrm{PV,S}} - C^{\mathrm{PV,P}}\left(P_{i,t}^{\mathrm{PV*}} - P_{i,t}^{\mathrm{PV}}\right) - C^{\mathrm{WT,P}}\left(P_{i,t}^{\mathrm{WT*}} - P_{i,t}^{\mathrm{WT}}\right), \tag{18}$$

which include generation revenue ($C^{\mathrm{WT,R}} P_{i,t}^{\mathrm{WT}}$, $C^{\mathrm{PV,R}} P_{i,t}^{\mathrm{PV}}$) and shared cost ($f_{i,t}^{\mathrm{WT,S}}, f_{i,t}^{\mathrm{PV,S}}$). $f_{i,t}^{\mathrm{WT,S}}, f_{i,t}^{\mathrm{PV,S}}$ are calculated as [7]

$$\begin{cases} f_{i,t}^{\mathrm{PV,S}} = C_t^{\mathrm{TN,D}} \bullet P_{i,t}^{\mathrm{PV,c}} \Big/ \sum_{g\in G} P_{g,t}^{\mathrm{c}} \\ f_{i,t}^{\mathrm{WT,S}} = C_t^{\mathrm{TN,D}} \bullet P_{i,t}^{\mathrm{WT,c}} \Big/ \sum_{g\in G} P_{g,t}^{\mathrm{c}}. \\ P_{i,t}^{\mathrm{PV,c}} = z_t^{\mathrm{PV,S}} P_{i,t}^{\mathrm{PV}} \\ P_{i,t}^{\mathrm{WT,c}} = z_t^{\mathrm{WT,S}} P_{i,t}^{\mathrm{WT}} \end{cases} \tag{19}$$

The compensation and shared cost of DPR service for directly connected generation can also be calculated similarly by referring to eqs. (14)-(17), and (19).

The constraints of WSTVPP operation are

$$P_{i,t}^{\mathrm{L}} = \sum_{u\in U_i} P_{i,u,t}^{\mathrm{TP}} + P_{i,t}^{\mathrm{WT}} + P_{i,t}^{\mathrm{PV}} + P_{i,t}^{\mathrm{UN}}, \tag{20}$$

$$\begin{cases} (1)P_{i,u,t}^{\mathrm{TP}} \in \left[\overline{P}_{i,u}^{\mathrm{TP}} - R_{i,u}^{\mathrm{TP,N}}\Delta t, \overline{P}_{i,u}^{\mathrm{TP}}\right] \\ -R_{i,u}^{\mathrm{TP,N}}\Delta t \le P_{i,u,t+1}^{\mathrm{TP}} - P_{i,u,t}^{\mathrm{TP}} \le R_{i,u}^{\mathrm{TP,N}}\Delta t \\ (2)P_{i,u,t}^{\mathrm{TP}} \in \left[\underline{P}_{i,u}^{\mathrm{TP,N}} - R_{i,u}^{\mathrm{TP,D}}\Delta t, \overline{P}_{i,u}^{\mathrm{TP}} - R_{i,u}^{\mathrm{TP,N}}\Delta t\right] \\ \frac{R_{i,u}^{\mathrm{TP,D}}}{R_{i,u}^{\mathrm{TP,N}}} P_{i,u,t+1}^{\mathrm{TP}} - P_{i,u,t}^{\mathrm{TP}} \le R_{i,u}^{\mathrm{TP,D}}\Delta t - \left(1 - \frac{R_{i,u}^{\mathrm{TP,D}}}{R_{i,u}^{\mathrm{TP,N}}}\right)\underline{P}_{i,u}^{\mathrm{TP,N}} \\ \frac{R_{i,u}^{\mathrm{TP,D}}}{R_{i,u}^{\mathrm{TP,N}}} P_{i,u,t}^{\mathrm{TP}} - P_{i,u,t+1}^{\mathrm{TP}} \le R_{i,u}^{\mathrm{TP,D}}\Delta t - \left(1 - \frac{R_{i,u}^{\mathrm{TP,D}}}{R_{i,u}^{\mathrm{TP,N}}}\right)\underline{P}_{i,u}^{\mathrm{TP,N}} \\ (3)P_{n,t}^{\mathrm{TP}} \in \left[\underline{P}_{i,u}^{\mathrm{TP,D}}, \underline{P}_{i,u}^{\mathrm{TP,N}} - R_{i,u}^{\mathrm{TP,D}}\Delta t\right] \\ -R_{i,u}^{\mathrm{TP,D}}\Delta t \le P_{i,u,t+1}^{\mathrm{TP}} - P_{i,u,t}^{\mathrm{TP}} \le R_{i,u}^{\mathrm{TP,D}}\Delta t \end{cases}, \tag{21}$$

$$\begin{cases} P_{i,u,t}^{\mathrm{TP}} = P_{i,u,t}^{\mathrm{TP,B}} + P_{i,u,t}^{\mathrm{TP,F}} \\ -\tau_{i,u}^{\mathrm{TP}} \overline{P}_{i,u}^{\mathrm{TP}} \Delta t \le P_{i,u,t}^{\mathrm{TP,F}} \le \tau_{i,u}^{\mathrm{TP}} \overline{P}_{i,u}^{\mathrm{TP}} \Delta t, \\ \underline{P}_{i,u}^{\mathrm{TP,D}} \le P_{i,u,t}^{\mathrm{TP}} \le \overline{P}_{i,u}^{\mathrm{TP}} \end{cases} \tag{22}$$

$$\begin{cases} \tau_i^{\mathrm{WT}} P_{i,t}^{\mathrm{WT}*} \le P_{i,t}^{\mathrm{WT}} \le P_{i,t}^{\mathrm{WT}*} \\ \tau_i^{\mathrm{PV}} P_{i,t}^{\mathrm{PV}*} \le P_{i,t}^{\mathrm{PV}} \le P_{i,t}^{\mathrm{PV}*} \end{cases}, \tag{23}$$

where (20) is the power balance constraint in WSTVPP. (21) is the nonconvex ladder-type ramping constraints of TPUs [8] in $\Delta t = 1/4\mathrm{h} = 15\mathrm{min}$. (22) is the output constraints and real-time flexibility reserve of TPUs. The real-time flexibility reserve of TPUs is used to correct the influence of uncertainties and power unbalanced between generation output and load demand. (23) is the output constraint of RES with utilization rates.

3. Federated Reinforcement Learning Framework

The multi-WSTVPP operation model in TN proposed above has nonlinear objective and non-convex constraints. Specifically, it comprises power flow and transmission constraints, operation cost of TPUs (12), share cost for DPR services (14)-(17), (19), and ramping constraint of TPUs (21). It is difficult for traditional optimization methods to efficiently handle this large-scale non-convex optimization problem to achieve both global and local cost-effective power sources scheduling. In this section, the multiple WSTVPPs operation is modeled as a Dec-POMDP, the designed action transformations are used for converting the output actions of the agent to the actual output of power sources in WSTVPP while reducing the unbalanced active power. Then the FRL framework is adopted to efficiently train the agent with privacy preservation.

3.1. Partially observable Markov decision process modeling

In this paper, to address the separate operation of TN and each WSTVPP, each WSTVPP can only access the local information and TSO dispatching instruction, e.g., the partial information in TN, to schedule its power sources. Therefore, it is reasonable to model the time series multi-WSTVPP operation process as a Dec-POMDP $\{\mathscr{I}, \mathscr{S}, \mathscr{O}, \mathscr{A}, \mathscr{R}, \mathscr{T}, \gamma\}$ [36] in the environment of TN with multiple WSTVPPs. Specifically, $i \in \mathscr{I}$ is the agent set, $\mathscr{S}$ is the global state, $o_i \in \mathscr{O}_{i\in\mathscr{I}}$ is the local observation set, $a_i \in \mathscr{A}_{i\in\mathscr{I}}$ is the action set, $r_i \in \mathscr{R}_{i\in\mathscr{I}}$ is the reward set, $\mathscr{T}$ is the stochastic state transition, γ is the discount factor. The specific elements of Dec-POMDP are:

1) Agent set: Each VPP controller is considered as an agent (i.e., WSTVPP control agent) to schedule its power sources.

2) Environment set: The environment is the TN OPF model and WSTVPP operation model in Sections II-B and C.

3) State and observation set: The global state is the combination of information in TN and WSTVPPs. The local observation of each WSTVPP is the local information of power sources and dispatch instruction of TSO:

$$o_{i,t} = \left\{ P_{i,t}^{\mathrm{WT}*}, \cdots, P_{i,t+h\Delta t}^{\mathrm{WT}*}, P_{i,t}^{\mathrm{PV}*}, \cdots, P_{i,t+h\Delta t}^{\mathrm{PV}*}, P_{i,t}^{\mathrm{L}}, \underline{P}_{i,t}^{\mathrm{NET}}, P_{i,u,t-1}^{\mathrm{TP}} (u \in U_i) \right\}. \tag{24}$$

It comprises short-term RES forecast for h consecutive time steps from t to $t + h\Delta t$, dispatch instruction, minimum net load at t based on dispatch instruction and RES forecast $\underline{P}_{i,t}^{\mathrm{NET}} = P_{i,t}^{\mathrm{L}} - P_{i,t}^{\mathrm{WT}*} - P_{i,t}^{\mathrm{PV}*}$, output of TPUs at $t-1$. The short-term forecast values account for the uncertainty of RESs with rolling adjustments made at each time step. The handling of these uncertainties will be detailed in the section on state transitions. Given the high accuracy of ultra-short-term forecasting [47], this paper directly uses $\mathrm{P}_{\mathrm{i,t}}^{\mathrm{WT}}*$ and $\mathrm{P}_{\mathrm{i,t}}^{\mathrm{PV}}*$ as the maximum available output of RESs at time t.

4) Action set: The action is set as the output range of power sources in each WSTVPP at t:

$$a_{i,t} = \left\{ a_{i,u,t}^{\mathrm{TP}} (u \in U_i), a_{i,t}^{\mathrm{WT}}, a_{i,t}^{\mathrm{PV}} \right\} \in [-1, 1]. \tag{25}$$

$a_{i,t}$ can be transformed to the actual output of power sources $P_{i,u,t}^{\mathrm{TP,B}} (u \in U_i)$, $P_{i,t}^{\mathrm{WT}}$, $P_{i,t}^{PV}$ at next time step based on their output constraints:

$$P_{i,t}^{(\bullet)} = 0.5\left(a_{i,t}^{(\bullet)} + 1\right)\left(\overline{a}_{i,t}^{(\bullet)} - \underline{a}_{i,t}^{(\bullet)}\right) + \underline{a}_{i,t}^{(\bullet)}, \tag{26}$$

where $(\bullet)$ represents a specific power source in WSTVPP (i.e., TPUs, WT, PV). For WT and PV, $\overline{a}_{i,t}^{\mathrm{WT}}$, $\overline{a}_{i,t}^{\mathrm{PV}}$ are $P_{i,t}^{\mathrm{WT}*}$, $P_{i,t}^{\mathrm{PV}*}$, $\underline{a}_{i,t}^{\mathrm{WT}}$, $\underline{a}_{i,t}^{\mathrm{PV}}$ are $\tau_i^{\mathrm{WT}} P_{i,t}^{\mathrm{WT}*}$, $\tau_i^{\mathrm{PV}} P_{i,t}^{\mathrm{PV}*}$. For TPUs, $\overline{a}_{i,u,t}^{\mathrm{TP}}$ and $\underline{a}_{i,u,t}^{\mathrm{TP}}$ can be determined by ramping constraints (21) and current net load ($P_{i,t}^{\mathrm{NET}} = P_{i,t}^{\mathrm{L}} - P_{i,t}^{\mathrm{WT}} - P_{i,t}^{\mathrm{PV}}$) considering compensation standard of DPR. For $P_{i,t}^{\mathrm{NET}} < \sum_{u\in U_i} \mu_1^{\mathrm{TP,D}} \overline{P}_{i,u}^{\mathrm{TP}}$, $\overline{a}_{i,u,t}^{\mathrm{TP}}$ and $\underline{a}_{i,u,t}^{\mathrm{TP}}$ are calculated based on the ramping constraints of TPU

$$\begin{cases} (1) P_{i,u,t}^{\mathrm{TP}} \in \left[\overline{P}_{i,u}^{\mathrm{TP}} - R_{i,u}^{\mathrm{TP,N}} \Delta t, \overline{P}_{i,u}^{\mathrm{TP}} \right] \\ \overline{a}_{i,u,t}^{\mathrm{TP}} = \min\left(\overline{P}_{i,u}^{\mathrm{TP}}, P_{i,u,t-1}^{\mathrm{TP}} + R_{i,u}^{\mathrm{TP,N}} \Delta t \right) \\ \underline{a}_{i,u,t}^{\mathrm{TP}} = P_{i,u,t-1}^{\mathrm{TP}} - R_{i,u}^{\mathrm{TP,N}} \Delta t \\ (2) P_{i,u,t}^{\mathrm{TP}} \in \left[\underline{P}_{i,u}^{\mathrm{TP,N}} - R_{i,u}^{\mathrm{TP,D}} \Delta t, \overline{P}_{i,u}^{\mathrm{TP}} - R_{i,u}^{\mathrm{TP,N}} \Delta t \right] \\ \overline{a}_{i,u,t}^{\mathrm{TP}} = min\left\{ \overline{P}_{i,u}^{\mathrm{TP}}, \frac{R_{i,u}^{\mathrm{TP,N}}}{R_{i,u}^{\mathrm{TP,D}}} \left[P_{i,u,t-1}^{\mathrm{TP}} + R_{i,u}^{\mathrm{TP,D}} \Delta t - \left(1 - \frac{R_{i,u}^{\mathrm{TP,D}}}{R_{i,u}^{\mathrm{TP,N}}} \right) \underline{P}_{i,u}^{\mathrm{TP,N}} \right] \right\}. \\ \underline{a}_{i,u,t}^{\mathrm{TP}} = max\left\{ \underline{P}_{i,u}^{\mathrm{TP,D}}, \frac{R_{i,u}^{\mathrm{TP,D}}}{R_{i,u}^{\mathrm{TP,N}}} P_{i,u,t}^{\mathrm{TP}} - R_{i,u}^{\mathrm{TP,D}} \Delta t - \left(1 - \frac{R_{i,u}^{\mathrm{TP,D}}}{R_{i,u}^{\mathrm{TP,N}}} \right) \underline{P}_{i,u}^{\mathrm{TPU,N}} \right\} \\ (3) P_{i,u,t}^{\mathrm{TP}} \in \left[\underline{P}_{i,u}^{\mathrm{TP,D}}, \underline{P}_{i,u}^{\mathrm{TP,N}} - R_{i,u}^{\mathrm{TP,D}} \Delta t \right] \\ \overline{a}_{i,u,t}^{\mathrm{TP}} = P_{i,u,t-1}^{\mathrm{TP}} + R_{i,u}^{\mathrm{TP,D}} \Delta t \\ \underline{a}_{i,u,t}^{\mathrm{TP}} = max\left(\underline{P}_{i,u}^{\mathrm{TP,D}}, P_{i,u,t-1}^{\mathrm{TP}} - R_{i,u}^{\mathrm{TP,D}} \Delta t \right) \end{cases} \tag{27}$$

For $P_{i,t}^{\mathrm{NET}} \ge \sum_{u\in U_i} \mu_1^{\mathrm{TP,D}} \overline{P}_{i,u}^{\mathrm{TP}}$, TPUs cannot obtain revenue from DPR service. In this case, $\overline{a}_{i,u,t}^{\mathrm{TP}}$ is calculated similarly as (27), $\underline{a}_{i,u,t}^{\mathrm{TP}}$ is modified as follow to help reduce DPR loss

$$\underline{a}_{i,u,t}^{\mathrm{TP}} = \min\left(P_{i,u,t-1}^{\mathrm{TP}} + R_{i,u}^{\mathrm{TP,D}} \Delta t, \underline{P}_{i,u}^{\mathrm{TP,N}} \right). \tag{28}$$

After transforming the actions of TPUs from $\mathrm{a}_{\mathrm{i,u,t}}^{\mathrm{TP}}$ to $P_{i,u,t}^{\mathrm{TP,B}}$, the real-time fine-tune $P_{i,u,t}^{\mathrm{TP,F}}$ in (22) is introduced to reduce the remaining $P_{i,t}^{\mathrm{UN}}$. The fine-tuning order of TPUs is determined by the economy of the TPUs in a WSTVPP. For $P_{i,t}^{\mathrm{UN}} > 0$, the power generation of WSTVPP is insufficient, in this case, TPU with better economic performance is prioritized for fine-tuning to increase their output until $P_{i,t}^{\mathrm{UN}} = 0$. For $P_{i,t}^{\mathrm{UN}} < 0$, the WSTVPP is overgeneration, TPU with lower economic performance is prioritized to decrease its output until $P_{i,t}^{\mathrm{UN}} = 0$.

5) State transition: The state transition $\mathscr{T}$ is defined as a stochastic transition function from state $\mathscr{S}_t$ to $\mathscr{S}_{t+1}$ after each agent takes action $\mathscr{A}_{i\in\mathscr{I}}$. The state transition processes are: (i) Each agent takes action $a_{i,t}$ through local observation $o_{i,t}$, transforms $a_{i,t}$ to actual output $P_{i,u,t}^{\mathrm{TP}} (u \in U_i)$, $P_{i,t}^{\mathrm{WT}}$, $P_{i,t}^{\mathrm{PV}}$. (ii) Each WSTVPP control agent calculates its DPR compensation revenue $\sum_{u\in U_i} f_{i,u,t}^{\mathrm{TP,R,D}}$ by eq. (11), total corrected output $P_{i,t}^{c}$ by eqs. (15), (16), (19), and feasible output range at next time step $\overline{P}_{g,t+1}$, $\underline{P}_{g,t+1}$, $\overline{Q}_{g,t+1}$, $\underline{Q}_{g,t+1}$ by eqs. (4)-(7). (iii) Each agent sends the information in (ii) to TSO. (iv) The TSO calculates the total DPR

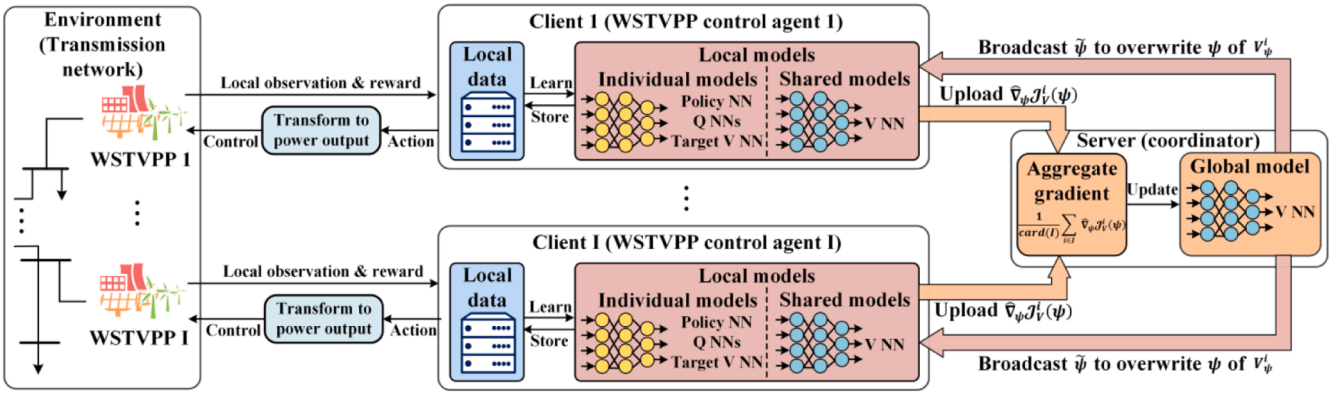


Fig. 2. FRL algorithm for the operation of multiple WSTVPPs.

compensation revenue in TN $C_t^{\mathrm{TN,D}}$ by eq. (17), total corrected output $\sum_{g\in G} P^c_{g,t}$, dispatching instruction of each grid-connected power source and each WSTVPP at the next time step $P_{i,t+1}$, $Q_{i,t+1}$ by OPF, and sends them to each agent. (v) Each agent calculates its share cost, reward $r_{i,t}$, and updates local observation $o_{i,t+1}$ for the next time step. Notably, the WT, PV in $o_{i,t}$ and load in TN have stochastic characteristics [37]. Gaussian noise is added to the RES forecast values and load values to represent this stochastics in the state transition process. The FRL framework can handle this stochastic and output correct action through training.

6) Reward: The reward of each agent is designed to maximize the objective of its corresponding WSTVPP and avoid voltage and power flow exceeding limits in TN at t

$$r_{i,t} = f^{\mathrm{VPP}}_{i,t} - b_{opf}\omega_{opf}, \tag{29}$$

where a large penalty ω_{opf} and a binary variable $b_{opf} = 1$ are introduced if the OPF is unsolvable to ensure the outputs of power sources are in allowable ranges for safe operation in TN.

3.2. Federated reinforcement learning algorithm design

In this paper, a designed FRL framework is adopted to train the multiple WSTVPP control agents efficiently considering global and local cost-effective operation with privacy preservation, as shown in Fig. 2. The FRL framework has a server as coordinator with a global NN model to help enhance the training efficiency and cooperation among agents, and several distributed clients with individual NN models and data for agents. Specifically, the server has a parameterized global state value NN (V network) $V^{\mathrm{G}}_{\bar{\psi}}(o_t)$. Each client deploys a corresponding WSTVPP control agent with local data and five parameterized individual local NNs: a policy NN $\pi^i_{\varnothing(i)}(a_{i,t}|o_{i,t})$, two state action value NNs (Q networks) $Q^i_{\theta 1(i)}(o_{i,t}, a_{i,t})$ and $Q^i_{\theta 2(i)}(o_{i,t}, a_{i,t})$, a V network $V^i_{\psi(i)}(o_{i,t})$ and a target V network $V^i_{\bar{\psi}(i)}(o_{i,t})$ with the same structure as the global $V^{\mathrm{G}}_{\bar{\psi}}(o_t)$. Each client also has a replay buffer D^i to store the local data $(o_{i,t}, a_{i,t}, r_{i,t}, o_{i,t+1}, d_{i,t})$ for each time step during training, where $d_{i,t}$ is a binary variable to record whether an episode is done (1 if $t = T$, else 0). At the client, each agent's policy NN $\pi^i_{\varnothing(i)}(a_{i,t}|o_{i,t})$ is trained to schedule the power sources in each WSTVPP based on the local observation for maximizing the accumulative discounted reward $\sum_{t=0}^{T}\gamma^t r_{i,t}$. The individual policy is optimized by updating iteratively with the help of other local NN models, local data, and global NN model. During the update process, only the gradients and parameters of the global NN model are exchanged between server and clients. Therefore, the performance of agents can be improved while preserving the private local models and data for each agent.

The policy update process is based on the widely used soft-actor-critic (SAC) algorithm [38] specially designed for FRL framework. For each agent at the client, firstly, local Q networks are updated by minimizing

$$\begin{aligned}\mathscr{J}^i_Q[\theta(i)] &= \mathbb{E}_{o_{i,t},a_{i,t}\sim D^i}\left\{\frac{1}{2}\left[Q^i_{\theta(i)}(o_{i,t},a_{i,t}) - \widehat{Q}^i(o_{i,t},a_{i,t})\right]^2\right\}, \theta(i)\\ &= \theta 1(i), \theta 2(i),\end{aligned} \tag{30}$$

$$\widehat{Q}^i(o_{i,t},a_{i,t}) = r_{i,t} + (1-d_{i,t})\gamma\mathbb{E}_{o_{t+1}\sim p}\left[V_{\bar{\psi}(i)}(o_{i,t+1})\right], \tag{31}$$

where p is a stochastic state transition probability. $Q^i_{\theta 1(i)}(o_{i,t}, a_{i,t})$ and $Q^i_{\theta 2(i)}(o_{i,t}, a_{i,t})$ can be updated by stochastic gradient descent (SGD) through the gradients $\widehat{\nabla}_{\theta 1(i)}\mathscr{J}^i_Q[\theta 1(i)]$ and $\widehat{\nabla}_{\theta 2(i)}\mathscr{J}^i_Q[\theta 2(i)]$ of (30). The gradients can be calculated by an open-source RL framework such as Pytorch [39].

Secondly, the local V network $V^i_{\psi(i)}(o_{i,t})$ is updated by the SGD and information exchange between clients and server. Each agent calculates the stochastic gradient $\widehat{\nabla}_{\psi(i)}\mathscr{J}^i_V[\psi(i)]$ of

$$\mathscr{J}^i_V[\psi(i)] = \mathbb{E}_{o_{i,t}\sim D^i}\left\{\frac{1}{2}\left(V^i_{\psi(i)}(o_{i,t}) - \mathbb{E}_{\hat{a}_{i,t}\,\pi^i_{\varnothing}}\left[\min\left(Q^i_{\theta 1(i)}(o_{i,t},\hat{a}_{i,t}), Q^i_{\theta 2(i)}(o_{i,t},\hat{a}_{i,t})\right) - \alpha log\,\pi^i_{\varnothing(i)}(\hat{a}_{i,t}|o_{i,t})\right]\right)^2\right\}, \tag{32}$$

where $\hat{a}_{i,t}$ is the stochastic evaluation of action. To obtain the evaluation $\hat{a}_{i,t}$, a noise vector sampled from some fixed distribution is added to the output action of current policy based on $o_{i,t}$ in D^i. α is a temperature parameter of entropy term.

In the FRL framework, to achieve global coordination and training efficiency improvement, $V^i_{\psi(i)}(o_{i,t})$ is updated with the help of server. the server aggregates the gradient $\widehat{\nabla}_{\psi(i)}\mathscr{J}^i_V[\psi(i)]$ of (38) from each agent to update the global $V^{\mathrm{G}}_{\bar{\psi}}(o_t)$ and sends the updated $\bar{\psi}$ to overwrite $\psi(i)$ in local $V^i_{\psi(i)}$ at each client. The details of $V^{\mathrm{G}}_{\bar{\psi}}(o_t)$ update process at server will be discussed below.

Thirdly, the local policy network $\pi^i_{\varnothing(i)}(a_{i,t}|o_{i,t})$ is updated using the SGD with gradient $\widehat{\nabla}_{\varnothing(i)}\mathscr{J}^i_\pi(\varnothing(i))$ by minimizing

$$\mathscr{J}^{i}_{\pi}[\varnothing(i)] = \mathbb{E}_{o_{i,t}\sim D^{i}}\Big\{ a log\pi_{\varnothing(i)}\big(\widehat{a}_{i,t}|o_{i,t}\big) - \Big[\min\Big(Q^{i}_{\theta 1(i)}\big(o_{i,t},\widehat{a}_{i,t}\big), Q^{i}_{\theta 2(i)}\big(o_{i,t},\widehat{a}_{i,t}\big)\Big) - V^{i}_{\psi(i)}\big(o_{i,t}\big)\Big]\Big\}, \quad (33)$$

where $\min(Q^{i}_{\theta 1(i)}(o_{i,t},\widehat{a}_{i,t}), Q^{i}_{\theta 2(i)}(o_{i,t},\widehat{a}_{i,t})) - V^{i}_{\psi(i)}(o_{i,t})$ is the advantage term to prevent the overestimation of Q [40].

Finally, the $\overline{\psi}(i)$ in local target V NN $V^{i}_{\overline{\psi}}(o_{i,t})$ is updated by using the exponentially moving average of $\psi(i)$ in $V^{i}_{\psi(i)}(o_{i,t})$.

At server, the global model $V^{G}_{\widetilde{\psi}}(o_t)$ is updated by SGD, the corresponding gradient $\widehat{\nabla}_{\widetilde{\psi}}\mathscr{J}^{G}_{V}(\widetilde{\psi})$ is calculated by averaging the gradients aggregated from each client

$$\widehat{\nabla}_{\widetilde{\psi}}\mathscr{J}^{G}_{V}(\widetilde{\psi}) = \frac{1}{card(I)}\sum_{i\in I}\widehat{\nabla}_{\psi}\mathscr{J}^{i}_{V}(\psi), \quad (34)$$

where $card(I)$ is the number of elements in I, which equivalents to the number of agents.

Algorithm 1 FRL Framework for multiple WSTVPPs Operation

Client i **Initializ** $V^{i}_{\psi(i)}(o_{i,t})$; $\pi^{i}_{\varnothing(i)}(a_{i,t}|o_{i,t})$, $V^{i}_{\overline{\psi}(i)}(o_{i,t})$, $Q^{i}_{\theta 1(i)}(o_{i,t},a_{i,t})$, $Q^{i}_{\theta 2(i)}(o_{i,t},a_{i,t})$ with $\lambda_{\varnothing}$, $\lambda_{\overline{\psi}}$, $\lambda_{\theta 1}$, $\lambda_{\theta 2}$; D^{i}.
Server Initialization: $V^{G}_{\widetilde{\psi}}(o_t)$ with the same structure as $V^{i}_{\psi(i)}(o_{i,t})$ and $\lambda_{\widetilde{\psi}}$, broadcasts $\widetilde{\psi}$ to overwrite ψ of each agent's $V^{i}_{\psi(i)}(o_{i,t})$. B, E, T.
For episode $e = 1$ to E **do**
Reset environment: randomly select a start time, get the initial actions and local observation for each agent.
For environment time step t = 1 to T **do**
At client, each WSTVPP control agent i **do in parallel**
Randomly take actions $a_{i,t}$ via local observation $o_{i,t}$ by individual local policy $\pi^{i}_{\varnothing(i)}$
Transform actions to $P^{\mathrm{TP,B}}_{i,u,t}(u \in U_i)$, $P^{\mathrm{WT}}_{i,t}$, $P^{\mathrm{PV}}_{i,t}$.
Fine-tune TPUs by $P^{\mathrm{TP,F}}_{i,u,t}$ to acquire $P^{\mathrm{TP}}_{i,u,t}$.
Calculate $\sum_{u\in U_i} f^{\mathrm{TP,R,D}}_{i,u,t}$ and $P^{\mathrm{c}}_{i,t}$, send them to TSO.
Calculate $\overline{P}_{i,t+1}$, $\underline{P}_{i,t+1}$, $\overline{Q}_{i,t+1}$, $\underline{Q}_{i,t+1}$, send them to TSO.
At TSO
Calculate $C^{\mathrm{TN,D}}_{t}$ and $\sum_{g\in G}P^{\mathrm{c}}_{g,t}$ in TN, send to each grid-connected power sources and WSTVPPs.
Determine $\overline{P}_{g,t+1}$, $\underline{P}_{g,t+1}$, $\overline{Q}_{g,t+1}$, $\underline{Q}_{g,t+1}$ of grid-connected power sources and WSTVPP by their feasible ranges.
Run OPF. **If** OPF is solvable **do:**
Send $b_{opf} = 0$, dispatching instructions $P_{g,t+1}$, $Q_{g,t+1}$ to WSTVPPs and grid-connected power sources.
Else do:
Send $b_{opf} = 1$ to WSTVPPs.
end if
At client, each WSTVPP control agent i **do in parallel**
Calculate $\sum_{u\in U_i} f^{\mathrm{TP,S}}_{i,u,t}$, $f^{\mathrm{WT,S}}_{i,t}$, $f^{\mathrm{PV,S}}_{i,t}$.
Calculate reward $r_{i,t}$, obtain done $d_{i,t}$.
If $b_{opf} = 0$ **do:** Update local observation to $o_{i,t+1}$.
If $b_{opf} = 1$ **do:** $o_{i,t+1} = o_{i,t}$.
Store $\{o_{i,t}, a_{i,t}, r_{i,t}, o_{i,t+1}, d_{i,t}\}$ to D^{i}.
If $b_{opf} = 1$ **do: Break.**
$o_{i,t} \leftarrow o_{i,t+1}$.
If the update condition is satisfied **do:**
(Every certain step and **there are sufficient data in** D^{i}**)**
At client, each agent i **do in parallel**
Sample B batch of buffer from D^{i}
Update $\theta 1(i)$ of $Q^{i}_{\theta 1(i)}$: $\theta 1 \leftarrow \theta 1 - \lambda_{\theta 1}\widehat{\nabla}_{\theta 1}\mathscr{J}^{i}_{Q}(\theta 1)$.
Update $\theta 2(i)$ of $Q^{i}_{\theta 2(i)}$: $\theta 2 \leftarrow \theta 2 - \lambda_{\theta 1}\widehat{\nabla}_{\theta 2}\mathscr{J}^{i}_{Q}(\theta 2)$.
Calculate gradient $\widehat{\nabla}_{\psi(i)}\mathscr{J}^{i}_{V}[\psi(i)]$ and send to server.
At server
Aggregate $\widehat{\nabla}_{\psi}\mathscr{J}^{i}_{V}(\psi)$ and calculate $\widehat{\nabla}_{\widetilde{\psi}}\mathscr{J}^{G}_{V}(\widetilde{\psi})$.
Update $\widetilde{\psi}$ of $V^{G}_{\widetilde{\psi}}$: $\widetilde{\psi} \leftarrow \widetilde{\psi} - \lambda_{\psi}\widehat{\nabla}_{\widetilde{\psi}}\mathscr{J}^{G}_{V}(\widetilde{\psi})$.
Send $\widetilde{\psi}$ to each client.
At client, each agent i **do in parallel**
Overwrite $\psi(i)$ of V^{i}_{ψ} with $\widetilde{\psi}$ from server.
Update $\varnothing(i)$ of $\pi^{i}_{\varnothing(i)}$: $\varnothing \leftarrow \varnothing - \lambda_{\varnothing}\widehat{\nabla}_{\varnothing}\mathscr{J}^{i}_{\pi}(\varnothing)$.
Update $\overline{\psi}(i)$ of $V^{i}_{\overline{\psi}(i)}$: $\overline{\psi} \leftarrow \lambda_{\overline{\psi}}\psi + (1-\lambda_{\overline{\psi}})\overline{\psi}$.
end if (one update process is done)
end for (one episode is done)
end for (The training is done)

Table 1
Basic Data of Power).

Number	Bus	TPUs $\overline{P}^{\mathrm{TP}}$	$\underline{P}^{\mathrm{TP,N}}$	$\underline{P}^{\mathrm{TP,D}}$	PV S^{PV}	WT S^{WT}
VPP 1	32	200 + 300	100 + 150	/ + 120	200	300
VPP 2	33	300 + 300	150 + 150	120 + 120	120	180
VPP 3	38	300 + 600	150 + 300	120 + 180	200	200
VPP 4	39	600 + 600	300 + 300	180 + 180	400	300
G1	30	1040	520	/	/	/
G2	34	508	254	/	/	/
G3	35	687	343.5	/	/	/
G4	36	580	290	/	/	/
G5	37	564	282	/	/	/
Slack bus	31	/	/	/	/	/

Sources In Test Transmission Network (MW

3.3. Training process

The overall training process is shown in Algorithm 1. Before the start of the first training episode, each client initializes its individual $\pi^{i}_{\varnothing(i)}(a_{i,t}|o_{i,t})$, $Q^{i}_{\theta 1(i)}(o_{i,t}, a_{i,t})$, $Q^{i}_{\theta 2(i)}(o_{i,t}, a_{i,t})$, $V^{i}_{\psi(i)}(o_{i,t})$ with the same learning rate $\lambda_{\varnothing}$, $\lambda_{\theta 1}$, $\lambda_{\theta 2}$, λ_{ψ}, as well as $V^{i}_{\overline{\psi}(i)}(o_{i,t})$. The server initializes $V^{G}_{\widetilde{\psi}}(o_t)$ with the same structure as $V^{i}_{\psi}(o_{i,t})$ and learning rate $\lambda_{\widetilde{\psi}}$, broadcasts its parameters $\widetilde{\psi}$ to overwrite ψ of each agent's $V^{i}_{\psi}(o_{i,t})$.

At each training episode e, the environment is reset first. The server randomly selects the starting time step and broadcasts it to all clients for agents' training data synchronization. The TSO determines $\underline{P}_{g,t}$, $\overline{P}_{g,t}$ in (5) based on the summation of TPUs' minimum/maximum active output and predicted RES active outputs. The predicted RES outputs in the training process are the history data added with stochastic noises at corresponding time steps. The $\underline{Q}_{g,t}$, $\overline{Q}_{g,t}$ in (7) can be determined accordingly by $\underline{P}_{g,t}$, $\overline{P}_{g,t}$. Then the TSO runs the OPF program to solve the initial dispatching instruction for each grid-connected power source and WSTVPP. The initial active power output of grid-connected power source is set as its corresponding initial dispatching instruction. The initial active power outputs of RES in each WSTVPP are set as their corresponding predicted outputs. The initial active outputs of TPUs in each WSTVPP are set to allocate the remaining dispatching requirement (minimal net load mentioned in Section III-A-3)) proportionally based on the capacity of the TPUs. The above process resets the environment, then each agent can acquire the initial observation through the local and TSO information. Meanwhile, the TSO can obtain the $\underline{P}_{g,t}$, $\overline{P}_{g,t}$, $\underline{Q}_{g,t}$, $\overline{Q}_{g,t}$ by eqs. (4)-(7) at the next time step. The training episode can be carried out.

After resetting the environment, at each environment time step, each agent at the client takes actions $a_{\mathrm{i,t}}$ according to the current policy, transforms $a_{\mathrm{i,t}}$ to actual output of power sources $P^{\mathrm{TP}}_{i,u,t}(\mathrm{u} \in \mathrm{U_i})$, $P^{\mathrm{WT}}_{i,t}$, $P^{\mathrm{PV}}_{i,t}$. Then the state transition considering RES and load uncertainties in Section-III-A-5) is conducted, each agent obtains the next $o_{i,t+1}$ and stores $(o_{i,t}, a_{i,t}, r_{i,t}, o_{i,t+1}, d_{i,t})$ to replay buffer D^{i}, the environment moves to the next step. The above state-action-reward-next state circle is repeated until $t = T$, one episode is done. The training process is terminated until the maximum episode E is reached.

The model update process is triggered at every certain environmental step with sufficient data in the replay buffer. During the update process, each agent at the client samples B batches of replay buffer. Using the batches, the local and global models can be updated through the methods in Section III-B. The update process is also summarized in Algorithm 1.

During the validation or application process, each agent only needs to utilize its trained local policy NN to schedule the power sources' output of the corresponding WSTVPP.

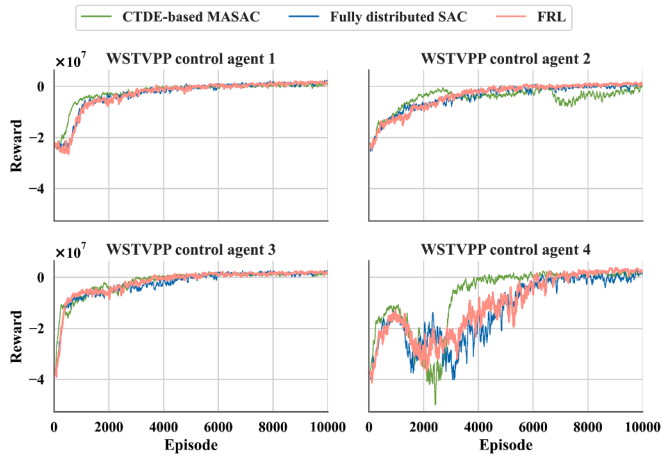
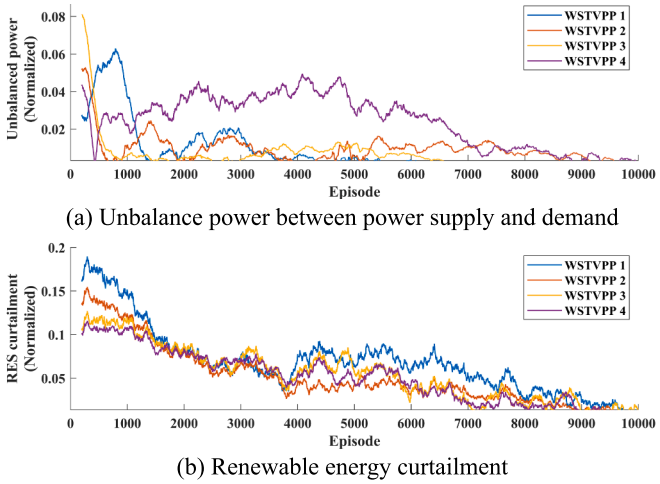


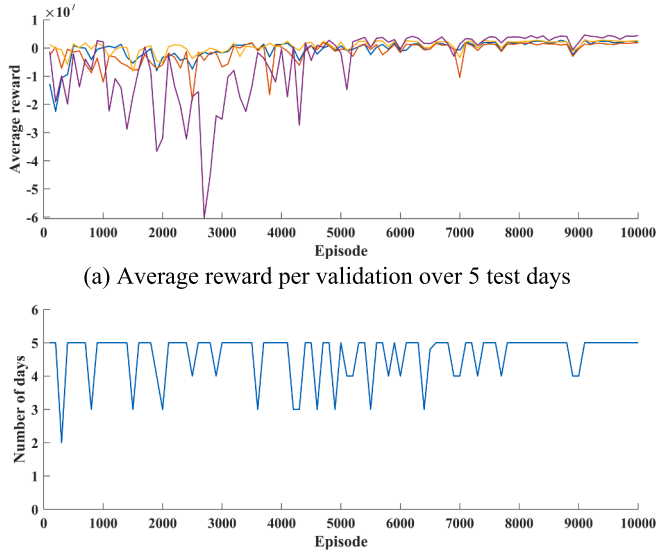
Fig. 3. Reward curves of each WSTVPP control agent during training (moving average).



(a) Unbalance power between supply and demand

(b) Renewable energy curtailment

Fig. 4. Some detailed information of each WSTVPP during training (moving average).



(a) Average reward per validation over 5 test days

(b) Count of test days achieving full-time step OPF convergence

Fig. 5. Validation of real-time training performance using the test data set (evaluated every 100 episodes).

4. Case Study

The proposed FRL-based multiple WSTVPPs coordinated operation framework is verified on a modified 39-bus TN [41], a widely used TN for studies. The test TN has 5 grid-connected TPUs with only NPR capability, 4 WSTVPPs (i.e., 4 agents, abbreviated as VPP) with WT, PV, and 2 retrofitted TPUs. The basic data of the power sources in TN is shown in Table 1. For the training data, the active and reactive load data are collected and modified based on [42] for 3 years. The WT and PV data are collected in [43] for 3 years. The test data set, distinct from the training data set, comprises data from representative days across various months of the year, accounting for approximately 15 % of the volume of the training data. The nonlinear AC OPF program adopts algorithm in Pandapower [44]. The voltage limit of the OPF in test TN is 0.94p.u. to 1.06p.u.. The ramping rate of all TPUs at NPR and DPR state are $0.9 \bullet S^{\mathrm{TP}}$/h and $0.6 \bullet S^{\mathrm{TP}}$/h. The parameters related to DPR auxiliary services are derived from reference [7]. $A_1^{\mathrm{TP,S}} = 0.7$, $A_2^{\mathrm{TP,S}} = 0.8$, $\mu_1^{\mathrm{TP,D}} = 0.5$, $\mu_2^{\mathrm{TP,D}} = 0.4$, $z_1^{\mathrm{TP,S}} = 1$, $z_2^{\mathrm{TP,S}} = 1.5$, $z_3^{\mathrm{TP,S}} = 2$. For winter season (November to next April), $k_t^{\mathrm{TP,D}} = 1$, $z_t^{\mathrm{PV,S}} = 2$, $z_t^{\mathrm{WT}} = 1.6$. For summer season (May to October), $k_t^{\mathrm{TP,D}} = 0.5$, $z_t^{\mathrm{PV}} = 1$, $z_t^{\mathrm{WT}} = 0.8$. The parameters for operation and costs are $w^{\mathrm{UN}} = 5 \times 10^4$, $\tau_i^{\mathrm{PV}} = \tau_i^{\mathrm{WT}} = 0.9$, $\tau_{i,u}^{\mathrm{TP}} = 0.4$, $C^{\mathrm{PV,P}} = C^{\mathrm{WT,P}} = 10^4$, $C_0^{\mathrm{TP,R}} = 375$, $C_1^{\mathrm{TP,R}} = 400$, $C_2^{\mathrm{TP,R}} = 1000$, $C^{\mathrm{PV,R}} = 740$, $C^{\mathrm{WT,R}} = 850$, $C^{\mathrm{coal}} = 685$, $C_m^{\mathrm{TP,E}} = 1.2$. All NN models in RL adopts Relu activation with size 64, $\lambda_{\varnothing} = \lambda_{\theta 1} = \lambda_{\theta 2} = \lambda_{\tilde{\psi}} = 4 \times 10^{-4}$, $\lambda_{\overline{\psi}} = 10^{-2}$, $\gamma = 0.99$, $\alpha = 1$, $B = 120$, $E = 10^4$, $T = 96$. The case study is run on a PC with Intel core i7-13700 K, 3.40 GHz and 32 GB of RAM.

4.1. Training performance

The proposed FRL algorithm is compared with other two RL approaches: (i) A fully distributed SAC algorithm that operates independently across 4 WSTVPP control agents without the aid of a central coordinator. Each agent only interacts with the TSO with its individual local NN models and data. (ii) A centralized training and decentralized execution-based multi-agent SAC (CTDE-based MASAC) approach that utilizes a central coordinator to aggregate data from all agents during the training phase for facilitating coordinated training efforts. Subsequently, the central coordinator shares the trained policy NN model with each agent for implementation. The learning rates of the above two RL approaches are aligned with that of the FRL algorithm.

The moving average reward curves of the three RL algorithms, based on 50 episodic rewards, for the 4 WSTVPP control agents are depicted in Fig. 3. The figure illustrates an initial trend of rewards starting from a notably negative value with similar converge speeds for the three algorithms. This initial phase reflects the agent's challenge in effectively managing the power balance constraint. Afterward, the rewards of FRL for all agents gradually converge to a stable level at about 10^6. While the fully distributed SAC for WSTVPP 2 and 4 converges to a relatively low reward. The CTDE-based MASAC demonstrates quicker improvement in the initial phases of training compared to both the proposed FRL and fully distributed SAC. However, the centralized training process, due to its handling of vast datasets simultaneously, experiences slower and less stable advancements during the latter stages of training. The final reward achieved by the CTDE-based MASAC is lower than that of the FRL method, with the rewards pertaining to WSTVPP 2 showing particular instability. Additionally, the centralized training process necessitates the complete sharing of agents' data and models with the coordinator, which compromises the privacy of individual agents. The computation burden of the central coordinator is also substantial. These results indicate that the proposed FRL algorithm is capable of learning an effective and stable policy coordinatively, while striking an optimal balance between preserving agent privacy and maintaining efficiency compared with the other two RL methods.

Table 2
Revenue Comparison of FRL and Optimization ($\times 10^6$ CNY).

	VPP 1	VPP 2	VPP 3	VPP 4
Summer day				
Proposed FRL	2.0319	2.3455	2.7821	4.9213
Optimization for each WSTVPP	2.1660	2.3504	2.8057	4.9563
Centralized optimization	1.9300	2.1480	1.6064	4.4956
Winter day				
Proposed FRL	5.5537	3.0990	5.3131	7.8196
Optimization for each WSTVPP	5.8318	3.1199	5.5212	8.0375
Centralized optimization	5.6477	2.8473	3.9075	7.3455

Table 3
Stability and Robustness Test of FRL Approach ($\times 10^6$ CNY).

		VPP 1	VPP 2	VPP 3	VPP 4
Max. load day	Proposed FRL	3.990	2.322	4.075	5.860
	Optimization	4.228	2.322	3.959	5.880
Max. load day, +5% load	Proposed FRL	4.199	2.731	4.160	6.167
	Optimization	4.366	2.794	4.115	6.208
Min. load day	Proposed FRL	4.011	2.165	4.257	5.882
	Optimization	4.231	2.203	4.072	6.035
Min. load day, −5% load	Proposed FRL	3.759	1.742	4.089	5.505
	Optimization	3.960	1.767	3.855	5.624
Max. RES day	Proposed FRL	3.650	2.142	3.566	5.928
	Optimization	3.831	2.271	3.695	5.982
Max. RES day, +5% RES	Proposed FRL	3.763	2.190	3.662	6.002
	Optimization	3.955	2.255	3.813	6.047
Min. RES day	Proposed FRL	2.036	1.793	2.585	4.255
	Optimization	2.169	1.798	2.577	4.267
Min. RES day, −5% RES	Proposed FRL	1.981	1.776	2.560	4.233
	Optimization	2.116	1.784	2.547	4.236

Some detailed information about each WSTVPP during the training phase is illustrated in Fig. 4. The figure reveals that the control agent of WSTVPP can engage in continuous learning through the proposed FRL algorithm. During the training process, the agent efficiently coordinates the power sources within the system in line with the dispatch instructions while reducing the RES curtailment. The above results demonstrate the effectiveness of the proposed method in facilitating collaborative training among agents with good improvement effects and resource utilization efficiency.

To evaluate the NN model's generalization capability, 5 test days are randomly selected from the test dataset to evaluate real-time performance every 100 training episodes. The results are presented in Fig. 5. Fig. 5 (a) shows a gradual stabilization of rewards for all agents over these test days. Furthermore, OPF achieves consistent convergence across the test days after about 9000 episodes, ensuring that voltage levels remained within the acceptable range of 0.94 to 1.06p.u., as shown in Fig. 5 (b). The results demonstrate that agent's policy NN model can achieve correct convergence and adapt to varying conditions through training with strong generalization capabilities.

4.2. Effectiveness validation for trained policy

To validate the training performance, the trained policy NN models of WSTVPP control agents by using the FRL algorithm are tested on a typical summer day and a typical winter day. The initial outputs of RESs are set to their predicted values. The initial outputs of TPUs are set to allocate $\underline{P}_{i,t}^{\mathrm{NET}}$ proportionally based on their $\overline{P}_{i,u}^{\mathrm{TP}}$. The start hour is 0:00. To enable consistent results, the stochastic noises of WT, PV and load are not considered during testing. For comparison, the mathematic model-based day-ahead optimization approaches are also conducted on these two typical days. These optimization approaches comprise independent optimization for each WSTVPP and centralized optimization for all power sources considering OPF in TN.

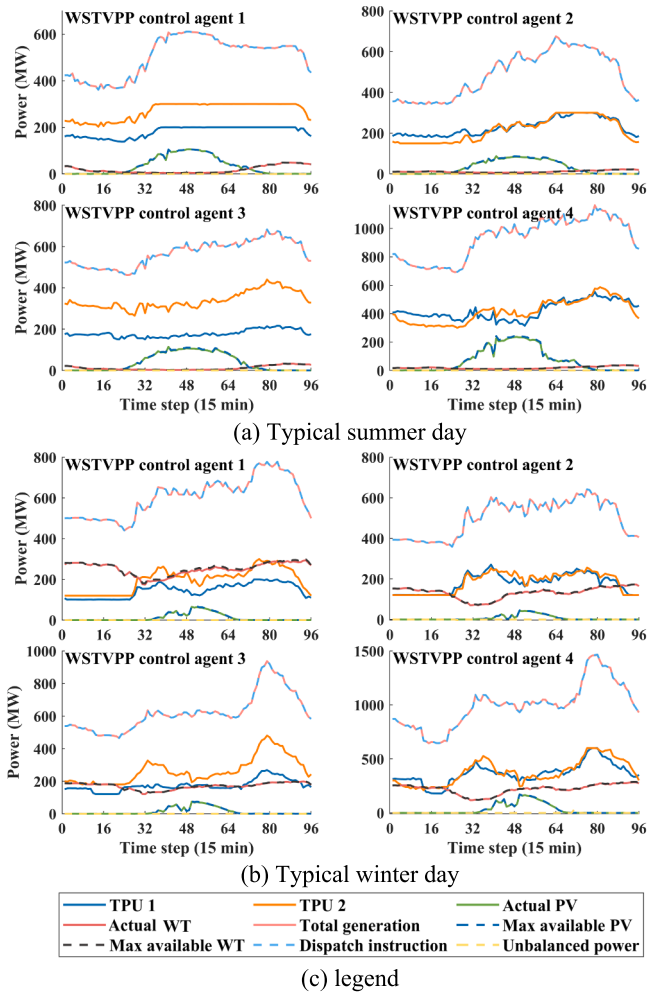
The optimization approach for each VPP adopts the following settings: (i) 4 WSTVPPs' operation models are considered separately as described in eqs. (8)-(22) without considering the shared cost constraints of DPR service eqs. (14)-(17), and (19). (ii) Due to the dynamic changes in the dispatch instructions of WSTVPPs in the proposed multiple power sources coordinated framework, the $P_{g,t}$, $Q_{g,t}$ used in the optimization program are derived through interactions between the WSTVPP control agents and TSO during the testing process of the proposed framework. (iii) The convex model is employed, $\Delta t = 1/2\mathrm{h} = 30\mathrm{min}$ is used to avoid nonconvex ramping constraints of TPUs in $\Delta t = 15\mathrm{min}$, the operation cost functions of TPUs are approximated using linearization, and the shared costs associated with DPR service are calculated after optimization is completed. (iv) The optimization problem is a mixed integer linear optimization, the Gurobi solver [45] is used to solve the optimization problem.

The centralized optimization approaches adopt the following settings: (i) The first approach, centralized optimization with only TN objective, structured as a nonlinear optimization, the objective is defined by eq. (1), with consideration of OPF constraints outlined in Section II-B, along with eqs. (20)-(22). The RES curtailment is not considered, $\Delta t = 1/2\mathrm{h} = 30\mathrm{min}$. All revenues and costs of power sources are calculated after optimization. (ii) The second approach, centralized optimization integrates both TN and WSTVPP objectives, formulated as a mixed integer nonlinear optimization, the objective is the summation of eq. (1) and eq. (8) for all WSTVPPs, with all constraints specified in Sections II-B and C. Similar to the first approach, the RES curtailment is not considered, $\Delta t = 1/2\mathrm{h} = 30\mathrm{min}$. (iii) The optimization models are programmed using the GAMS software [46] and are solved by the commercial solvers.

It should be noted that the comparison between the optimization and the proposed FRL-based framework is not entirely under identical conditions. Specifically, the scheduling time scales for FRL and optimization methods are different. The optimization for each WSTVPP relies on dispatch instructions derived from FRL outcomes, and its objectives exclude considerations of shared costs. Additionally, the goals of centralized optimization methods do not fully account for the multi-objective nature of WSTVPPs. Consequently, the results should be considered as indicative rather than definitive.

The revenues of each WSTVPP obtained by FRL-based trained policy and optimization approaches on the two typical days are presented in Table 2. The table provides clear evidence that the revenues obtained through the FRL exhibit only a negligible 0.4 % to 6 % deviation when compared to the results obtained through optimization for each WSTVPP. The error on the winter day is greater than on the summer day. Specifically, WSTVPP 1 exhibits a relatively larger error, while the errors for the other WSTVPPs typically hover around 1 %. The results presented above indicate that the proposed FRL approach is capable of training agents to achieve economical scheduling of power sources within WSTVPP, while concurrently fulfilling the dispatch requirements of the TN. Meanwhile, these tests highlight that optimization is a one-time calculation, lacking the capacity to adapt to real-time dynamics in TN. Conversely, the FRL method excels in adjusting power source outputs within the dynamic TN operating environment without relying on predefined scheduling plans.

In centralized optimization, scheduling multiple power sources within TN requires the TSO to access the detailed model of each power source. This requirement cannot protect the privacy of power source operators. Additionally, centralized optimization poses a large-scale, nonconvex problem characterized by substantial computational demands and high complexity. These attributes make it challenging to guarantee global optimal convergence within the constraints of a short-time scale multiple power sources scheduling. As presented in Table 2, the revenue of each WSTVPP obtained from centralized optimization, which focuses solely on TN objective, are lower than those of the proposed FRL method. This is because prioritizing only the TN's objective does not facilitate the TPU in maximizing benefits from DPR auxiliary services. Centralized optimization fails to find an optimal solution when

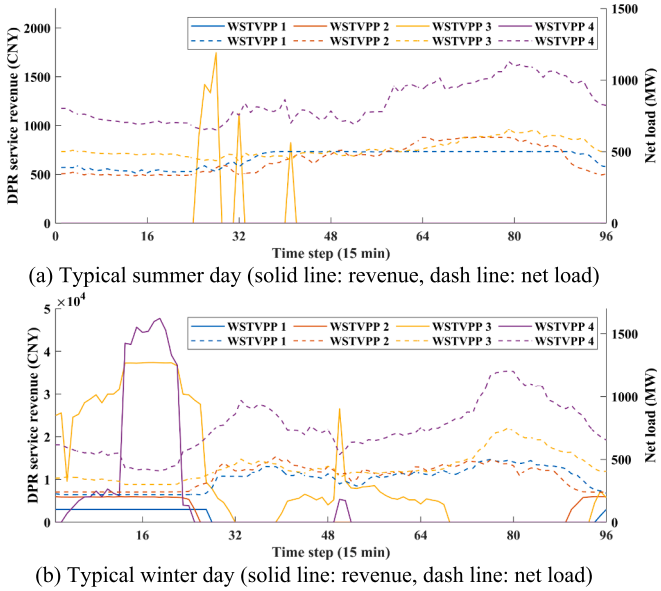


(a) Typical summer day

(b) Typical winter day

(c) legend

Fig. 6. Output of power sources obtained from FRL approach.



(a) Typical summer day (solid line: revenue, dash line: net load)

(b) Typical winter day (solid line: revenue, dash line: net load)

Fig. 7. DPR service revenues of TPUs at each time step.

attempting to balance objectives for both TN and WSTVPPs under nonconvex constraints by several commercial solvers such as SBB, CONOPT, and BARON in GAMS. This suggests that the comprehensive and complex multi-power operation model developed in this study is not suitable for traditional centralized optimization techniques.

To further assess the stability and robustness of the trained agents by the proposed FRL algorithm, days with the maximum and minimum load, as well as days with the highest and lowest RES output in the TN, are selected for testing. Moreover, future load growth and variability are also incorporated into these test scenarios. The results are shown in Table 3 and compared with the optimization approach for each VPP.

The table shows that under scenarios of maximum and minimum load or RES output, the discrepancy between revenues derived from the proposed FRL algorithm in this study and those obtained through the optimization method remains around a negligible 1 %. The optimization method's lack of consideration for DPR service cost-sharing results in the FRL approach achieving some higher revenues, as observed in the revenues of WSTVPP 3 on days of maximum and minimum loads. The above analysis highlights the adeptness of the FRL approach in handling various scenarios, addressing potential changes in load patterns and the variability of RES, thereby demonstrating its robustness. Moreover, for facing the potential increase in both load and renewable energy in the future, the proposed FRL algorithm can offer a direct and time-efficient strategy for ongoing agent training to adeptly meet upcoming requirements.

In summary, the proposed FRL-based multiple power sources coordinated framework is an effective solution for practical applications. It not only addresses privacy concerns but also provides the flexibility needed to adapt to real-time changes and large-scale nonconvex optimization across different scenarios in the TN with high overall revenue.

4.3. Detailed analysis of power sources scheduling

The detailed power sources scheduling obtained from the proposed FRL-based multiple power sources coordinated framework for every 15 min on the two typical days are shown in Fig. 6. The DPR service revenues of TPUs in each WSTVPP at each time step are shown in Fig. 7. From Fig. 6 (a)-(b), it is evident that the trained agents can effectively schedule TPUs and RESs in WSTVPPs to achieve real-time power balance between generation and dispatch instruction considering RES and load fluctuations. Although the TPUs are constrained by the maximum ramping rate in a short time scale, the trained agent can schedule the TPUs step by step based on local observations and different time periods to adapt to the changes in load and RES in the future with a longer time scale. For example, TPUs can increase their output ahead of peak load demand, typically occurring between 19:00 and 22:00 on winter days. Besides, the trained agents can also realize high-RES utilization and schedule TPUs to engage DPR service for more revenues. The details will be discussed below.

On the summer day depicted in Fig. 6 (a), the dispatch instruction for each WSTVPP remains relatively stable from 8:00 to 24:00. The PV output is high at noon, while the WT output is low throughout the day, resulting in a relatively modest total RES output. Besides, for the summer season, the revenue of DPR services is relatively low. Therefore, TPUs rarely engage in DPR service to mitigate additional losses in the DPR state. As depicted in Fig. 7 (a), only WSTVPP 3 engaged in the DPR service, specifically during the early morning hours of 6:00–8:00 and again at 10:00, when dispatch instruction is at its lowest.

Table 4
Renewable Energy Utilization of the Two Typical Days (%).

	WSTVPP 1	WSTVPP 2	WSTVPP 3	WSTVPP 4
Summer				
PV	99.82	99.81	99.69	99.65
WT	100	100	100	99.98
Winter				
PV	99.77	99.72	99.93	99.87
WT	99.88	99.90	100	100

Table 5
Total Revenue and Shared cost of DPR Service on the Winter Day ($\times 10^5$ CNY).

	WSTVPP 1	WSTVPP 2	WSTVPP 3	WSTVPP 4
Revenue	0.86	1.81	10.45	5.09
Shared cost	2.37	1.35	1.51	2.51

On the winter day illustrated in Fig. 6 (b), the dispatch instruction is relatively high from 18:00 to 24:00. The PV output is relatively low, whereas WT output is consistently high throughout the day, leading to the total RES output being higher than that of the summer day. This contributes to a reduced net load in comparison to the summer, particularly evident during nighttime. As demonstrated in Fig. 7 (b), the TPUs proactively reduce their outputs and engage in DPR service during times of low net load from 0:00 to 6:00, and also during midday when RES output is substantial. Revenue from DPR services is significantly higher in the winter, enhancing the incentive for TPUs to participate. The potential for DPR auxiliary service revenue within a WSTVPP is directly linked to the DPR capability of its TPUs. In particular, WSTVPPs 3 and 4, which are equipped TPUs possessing enhanced DPR capabilities (such as 600 MW TPU), are poised to achieve higher DPR service revenues in comparison to other WSTVPPs, especially during times of low net load in the TN from 0:00 to 7:00.

Table 4 shows the RES utilization rate. According to the table, the utilization rate of WT and PV over the two typical days is either close to or precisely 100 %. Specifically, the utilization rate of WT is slightly higher than that of PV. This discrepancy arises from the significant output changes of PV during the daytime. Additionally, the unit generation revenue of WT is higher than that of PV, leading WSTVPP to give priority to ensuring WT utilization. In WSTVPPs 1 and 2, the proportion of PV capacity in the RES is relatively small in comparison to WT, whereas WSTVPPs 3 and 4 exhibit the reverse scenario. Consequently, during significant PV output variations on summer days, WSTVPPs 1 and 2 achieve a higher PV utilization rate compared to WSTVPPs 3 and 4. Conversely, on winter days with substantial WT output, WSTVPPs 3 and 4, which have a smaller WT capacity and a larger thermal power capacity, can offer more flexibility to Integrate WT effectively.

Table 5 presents the total DPR revenue and shared cost of each WSTVPP on the winter day. The table shows that WSTVPPs with a wider output regulation range TPUs can gain more benefits from DPR services. Among them, WSTVPP 3 has the highest DPR service revenue due to the relatively low power demand of its dispatch instructions throughout the day. The DPR service revenue of WSTVPP 2, 3, and 4 surpasses their respective shared costs. These results indicate that aggregating high-performance flexible retrofitted TPUs with RES as a WSTVPP can achieve more overall revenue when participating in DPR services.

5. Conclusion

In this paper, a hybrid approach that combines data-driven FRL with the model-based method for multiple power sources coordinated operation in a wind-solar-thermal power network is proposed. The non-convex ramping constraint of TPUs, DPR service revenue and shared cost, as well as nonlinear cost functions, are considered in the multiple power sources operation model. By aggregating the megawatt-level grid-connected TPUs and RES plants at the same high voltage bus as the WSTVPP, and decomposing the multiple power sources scheduling model into the power network OPF model and WSTVPPs operation model, the privacy of the power sources operators can be preserved, the computational complexity can be reduced. The multiple WSTVPPs operation is modeled as Dec-POMDP and a customized FRL algorithm is adopted to train the WSTVPP control agents. Through a case study conducted in a 39-bus TN, the performance of the proposed FRL framework is validated. The trained agent can achieve economic operation and obtain more DPR service revenue for each WSTVPP considering privacy preservation.

CRediT authorship contribution statement

Yao Zou: Writing – original draft, Software, Data curation. **Qianggang Wang:** Writing – review & editing, Funding acquisition. **Qinqin Xia:** Software, Data curation. **Yuan Chi:** Writing – review & editing, Methodology. **Chao Lei:** Methodology. **Niancheng Zhou:** Writing – review & editing, 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

Data will be made available on request.

Acknowledgments

This work is supported by National Key Research and Development Program of China under Grant 2023YFB2405900.

References

[1] Hou Q, Du E, Zhang N, et al. Impact of High Renewable Penetration on the Power System Operation Mode: A Data-Driven Approach. IEEE Trans Power Syst 2020;35(1):731–41.
[2] Wang Y, Lou S, Wu Y, et al. Flexible Operation of Retrofitted Coal-Fired Power Plants to Reduce Wind Curtailment Considering Thermal Energy Storage. IEEE Trans Power Syst 2020;35(2):1178–87.
[3] Zhang J, Wang Y, Zhou G, et al. Integrating physical and data-driven system frequency response modelling for wind-PV-thermal power systems. IEEE Trans Power Syst 2023. https://doi.org/10.1109/TPWRS.2023.3242832.
[4] Rouzbahani HM, Karimipour H, Lei L. A review on virtual power plant for energy management. Sustain Energy Technol Assess 2021;14:101370.
[5] International Renewable Energy Agency (2019). Flexibility in conventional power plants. https://www.irena.org/-/media/Files/IRENA/Agency/Publication/2019/Sep/IRENA_Flexibility_in_CPPs_2019.pdf?la=en&hash=AF60106EA083E492638D8FA9ADF7FD099259F5A1.
[6] Agora Energiewende (2017). Flexibility in thermal power plants – With a focus on existing coal-fired power plant. https://www.agoraenergiewende.de/fleadmin2/Projekte/2017/Flexibility_in_thermal_plants/115_flexibility-report-WEB.pdf.
[7] Northeast China Energy Regulatory Bureau of National Energy Administration. (2020, Dec.). Operating rules of northeast electric power auxiliary service market. https://dbj.nea.gov.cn/xxgk/zcfg/202310/t20231011_147196.html.
[8] Yang L, Zhou N, Zhou G, et al. An Accurate ladder-type ramp rate constraint derived from field test data for thermal power unit with deep peak regulation. IEEE Trans Power Syst 2024;39(1):1408–20.
[9] Fusco A, Gioffrè D, Castelli AF, et al. A multi-stage stochastic programming model for the unit commitment of conventional and virtual power plants bidding in the day-ahead and ancillary services markets. Appl Energy 2023;336:120739.
[10] Yang L, Zhou N, Zhou G, et al. Day-ahead Optimal Dispatch Model for Coupled System Considering Ladder-type Ramping Rate and Flexible Spinning Reserve of Thermal Power Units. J Mod Power Syst Clean Energy 2022;10(6):1482–93.
[11] Wang Y, Lou S, Wu Y, et al. Flexible operation of retrofitted coal fired power plants to reduce wind curtailment considering thermal energy storage. IEEE Trans Power Syst 2020;35(2):1178–87.
[12] Khoshjahan M, Dehghanian P, Moeini-Aghtaie M, et al. Harnessing ramp capability of spinning reserve services for enhanced power grid flexibility. IEEE Trans Ind Appl 2019;55(6):7103–12.
[13] Jiang T, Shen Z, Jin X, et al. Solution to Coordination of Transmission and Distribution for Renewable Energy Integration into Power Grids: An Integrated Flexibility Market. CSEE J Power Energy Syst 2023;9(2):444–58.
[14] Oskouei MZ, Mohammadi-Ivatloo B, Abapour M, et al. Privacy-preserving mechanism for collaborative operation of high-renewable power systems and industrial energy hubs. Appl Energy 2021;283:116338.
[15] Véliz C, Grunewald P. Protecting data privacy is key to a smart energy future. Nat Energy 2018;3(9):702–4.
[16] Naughton J, Wang H, Cantoni M, et al. Co-Optimizing Virtual Power Plant Services Under Uncertainty: A Robust Scheduling and Receding Horizon Dispatch Approach. IEEE Trans Power Syst 2021;36(5):3960–72.
[17] Yan X, Gao C, Song M, Chen T, Ding J, Guo M, et al. An IGDT-based Day-ahead Co-optimization of Energy and Reserve in a VPP Considering Multiple Uncertainties. IEEE Trans Ind Appl 2022;58(3):4037–49.
[18] Yan X, Gao C, Ming H, Abbes D. Optimal scheduling strategy and benefit allocation of multiple virtual power plants based on general nash bargaining theory. Int J Electr Power Energy Sys 2023;152:109218.

[19] Fan S, Liu J, Wu Q, et al. Optimal coordination of virtual power plant with photovoltaics and electric vehicles: A temporally coupled distributed online algorithm. Appl Energy 2020;277:115583.
[20] Wu C, Gu W, Zhou S, et al. Coordinated Optimal Power Flow for Integrated Active Distribution Network and Virtual Power Plants Using Decentralized Algorithm. IEEE Trans Power Syst 2021;36(4):3541–51.
[21] Gough M, Santos SF, Almeida A, et al. Blockchain-Based Transactive Energy Framework for Connected Virtual Power Plants. IEEE Trans Ind Appl 2022;58(1):986–95.
[22] Walters DC, Sheble GB. Genetic algorithm solution of economic dispatch with valve point loading. IEEE Trans Power Syst 1993;8(3):1325–32.
[23] Chen X, Qu G, Tang Y, Low S, et al. Reinforcement learning for selective key applications in power systems: Recent advances and future challenges. IEEE Trans Smart Grid 2022;13(4):2935–58.
[24] Yi Z, Xu Y, Wang X, et al. An Improved Two-Stage Deep Reinforcement Learning Approach for Regulation Service Disaggregation in a Virtual Power Plant. IEEE Trans Smart Grid 2022;13(4):2844–58.
[25] Ochoa T, Gil E, Angulo A, et al. Multi-agent deep reinforcement learning for efficient multi-timescale bidding of a hybrid power plant in day-ahead and real-time markets. Appl Energy 2022;317:119067.
[26] Liu X, Li S, Zhu J. Optimal Coordination for Multiple Network-Constrained VPPs via Multi-Agent Deep Reinforcement Learning. IEEE Trans Smart Grid 2023;14(4):3016–31.
[27] Ding L, Lin Z, Shi X, Yan G. Target-Value-Competition-Based Multi-Agent Deep Reinforcement Learning Algorithm for Distributed Nonconvex Economic Dispatch. IEEE Trans Power Syst 2023;38(1):204–17.
[28] Qi J, Zhou Q, Lei L, et al. Federated reinforcement learning: techniques, applications, and open challenges. arXiv preprint 2021;arXiv:2108.11887.
[29] Li Y, He S, Li Y, et al. Federated multiagent deep reinforcement learning approach via physics-informed reward for multimicrogrid energy management. IEEE Trans Neural Netw Learn Syst 2023. https://doi.org/10.1109/TNNLS.2022.3232630.
[30] Qiu D, Xue J, Zhang T, et al. Federated reinforcement learning for smart building joint peer-to-peer energy and carbon allowance trading. Appl Energy 2023;333:120526.
[31] Feng B, Liu Z, Huang G, et al. Robust federated deep reinforcement learning for optimal control in multiple virtual power plants with electric vehicles. Appl Energy 2023;349:121615.
[32] Liu H, Wu W. Federated Reinforcement Learning for Decentralized Voltage Control in Distribution Networks. IEEE Trans Smart Grid 2022;13(5):3840–3.
[33] Zhao Q, Liao W, Wang S, et al. Robust Voltage Control Considering Uncertainties of Renewable Energies and Loads via Improved Generative Adversarial Network. J Mod Power Syst Clean Energy 2020;8(6):1104–14.
[34] Adibi MM, Milanicz DP. Reactive capability limitation of synchronous machines. IEEE Trans on Power Syst 1994;9(1):29–40.
[35] Šepetanc K, Pandžić H. Convex Polar Second-Order Taylor Approximation of AC Power Flows: A Unit Commitment Study. IEEE Trans Power Syst 2021;36(4):3585–94.
[36] Garrabé É, Russo G. Probabilistic design of optimal sequential decision-making algorithms in learning and control. Annu Rev Control 2022;54:81–102.
[37] Zou Y, Wang Q, Hu B, et al. Hierarchical evaluation framework for coupling effect enhancement of renewable energy and thermal power coupling generation system. Int J Electr Power Energy Sys 2022;146:108717.
[38] Haarnoja T, Zhou A, Abbeel P, et al. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: in international conference on machine learning; 2018. p. 1861–70.
[39] Pytorch. https://pytorch.org/.
[40] Zhang B, Hu W, Cao D, et al. Soft actor-critic –based multi-objective optimized energy conversion and management strategy for integrated energy systems with renewable energy. Energy Convers Manag 2021;243:114381.
[41] Hiskens I. IEEE PES task force on benchmark systems for stability controls. 2013. http://www1.sel.eesc.usp.br/ieee/IEEE39/New_England_Reduced_Model_(39_bus_system)_MATLAB_study_report.pdf.
[42] Transparency on Grid Data. https://www.elia.be/en/grid-data/.
[43] Wang J, Xu W, Gu Y, et al. Multi-agent reinforcement learning for active voltage control on power distribution networks. Adv Neural Inf Process Syst 2021;34:3271–84.
[44] Pandapower. https://pandapower.readthedocs.io/en/v2.6.0/about.html.
[45] Gurobi. https://www.gurobi.com.
[46] GAMS. https://www.gams.com.
[47] Wang H, Lei Z, Zhang X, et al. A review of deep learning for renewable energy forecasting. Energy Conversion and Management 2019;198:111799.