batteries

MDPI

Article

Deep Reinforcement Learning-Based Method for Joint Optimization of Mobile Energy Storage Systems and Power Grid with High Renewable Energy Sources

Yongkang Ding [1], Xinjiang Chen [1] and Jianxiao Wang [2,*]

[1] Department of Industrial Engineering and Management, Peking University, Beijing 100871, China
[2] National Engineering Laboratory for Big Data Analysis and Applications, Peking University, Beijing 100871, China
* Correspondence: wang-jx@pku.edu.cn

Abstract: The joint optimization of power systems, mobile energy storage systems (MESSs), and renewable energy involves complex constraints and numerous decision variables, and it is difficult to achieve optimization quickly through the use of commercial solvers, such as Gurobi and Cplex. To address this challenge, we present an effective joint optimization approach for MESSs and power grids that consider various renewable energy sources, including wind power (WP), photovoltaic (PV) power, and hydropower. The integration of MESSs could alleviate congestion, minimize renewable energy waste, fulfill unexpected energy demands, and lower the operational costs for power networks. To model the entire system, a mixed-integer programming (MIP) model was proposed that considered both the MESSs and the power grid, with the goal of minimizing costs. Furthermore, this research proposed a highly efficient deep reinforcement learning (DRL)-based method to optimize route selection and charging/discharging operations. The efficacy of the proposed method was demonstrated through many numerical simulations.

Keywords: renewable energy; battery energy storage system; machine learning; deep reinforcement learning; data-driven optimization; cost minimization

Citation: Ding, Y.; Chen, X.; Wang, J. Deep Reinforcement Learning-Based Method for Joint Optimization of Mobile Energy Storage Systems and Power Grid with High Renewable Energy Sources. *Batteries* **2023**, *9*, 219. https://doi.org/10.3390/batteries9040219

Academic Editors: Rodolfo Dufo-López and Pascal Venet

Received: 9 February 2023
Revised: 19 March 2023
Accepted: 28 March 2023
Published: 5 April 2023

1. Introduction

Renewable energy sources such as wind, water, and solar energy have considerable capacity to lower energy costs and carbon emissions while playing a crucial role in creating low-carbon and sustainable energy systems, as highlighted in [1–3]. However, the intermittency and variability of renewable energy pose a great challenge to the safe and economic operation of power systems in regions with an abundance of renewable energy potential [4,5]. To address this challenge, energy storage systems (ESSs) are a promising technology for the integration of renewable energy and the reduction in renewable curtailment.

Stationary energy storage systems (SESSs), the most conventional application mode of BES, have enabled energy conversion and capacity sharing during a limited time period [6,7]. Specifically, given the predicted renewable energy generation curve, SESSs, renewable energy, and thermal units were scheduled collaboratively to integrate renewable energy. However, SESS was inflexible, as it relies on large-capacity and long-distance transmission lines. Meanwhile, the utilization efficiency of SESSs also depended on its location, as it was difficult for the owners of a SESS to determine the optimal location because numerous constraints had to be considered, such as political, economic, social, and geographical factors [8–10]. Therefore, mobile energy storage systems (MESSs), as an ongoing development application mode of BES that couples energy and transportation systems, incorporate vehicles (e.g., electric vehicles (EVs), trains, ships, etc.), batteries,

power converters, and transformers. As compared to SESSs, MESSs enable energy conversion and the capacity sharing of energy storage over longer time periods through their transportation network and considerably improve the utilization rate of battery assets and the emergency dispatching capability of the power system. MESSs had more flexible deployment capabilities and had a more remarkable capacity in energy arbitrage [11,12], renewable energy integration [13–15], peak-shaving [16,17], grid-congestion relief [18,19], and grid-investment deferral [20–22].

The authors of [11] proposed the concept of a utility-scale MESS, which incorporated electric trucks, energy storage, and energy conversion systems; constructed an optimization model involving charge scheduling and route planning of MESSs; and used the proposed model in the energy arbitrage of a power grid in California. The results revealed that, as compared to the SESS, the MESS had considerable benefits over a life-cycle time period. In [13], the researchers integrated MESS into a large-scale power distribution system in order to integrate renewable energy in remote areas and proposed a particle-swarm optimization to streamline a scheduling model of MESSs. The results showed that the proposed models and algorithm could considerably reduce the operating cost of the power system. The authors of [14] developed a decision framework that used MESS to improve system resilience for the emergency dispatching of the power distribution system, and they verified that the proposed decision framework could effectively improve system response-and-recovery speed in the emergency scenarios. In [18], the researchers integrated large-scale EVs into the electric energy distribution system and proposed a mixed-integer linear programming (MILP) model to optimize the charging schemes of EVs. The results showed that the proposed model could substantially alleviate grid congestion and improve the performance of EVs and grid benefits. The authors of [20] developed a collaborative planning model for SESSs and MESSs to reduce the capital cost of transmission lines, energy storage systems, and the operating costs of power systems. Meanwhile, ref. [20] used the proposed model to determine the optimal region for SESS and MESS implementation for China's northwestern grid. The aging issue of mobile energy storage devices has also attracted attention. Reference [23] analyzed the influence mechanism of grid-connected operation on the life degradation of lithium-ion battery energy storage systems and established a cost accounting model for frequency regulation, considering the impact of battery life degradation. Reference [24] studied the frequency regulation problem of a power grid model that includes loads, traditional generators, and multiple electric vehicles, aiming to minimize the degradation of battery devices. Their proposed strategies demonstrated high effectiveness under actual operating conditions. In this paper, in order to emphasize the modeling of mobile energy storage systems and their joint optimization with the power grid, we have simplified the description of this part.

Recently, deep reinforcement learning (DRL) techniques have seen growing applications in battery management [25,26], grid management, and micro-grid management [27,28]. The authors of [25] presented a DRL-based method for a battery system in EVs, which consisted of a high-power battery pack in order to reduce energy loss and enhance thermal safety. The proposed strategy demonstrated advantages in terms of reducing calculation time and energy consumption. In [26], the researchers proposed a joint optimization framework of DRL and binary integer programming for the battery swapping–charging systems for EVs, which had favorable performance in a real-time demonstration, as well as improved computational efficiency and privacy preservation. A multi-agent system based on an RL method was applied to manage a stand-alone micro-grid including a power-production unit, a power-consumption unit, and a power-storage unit in [27]. The authors of [28] proposed an optimal control strategy based DRL, which had asynchronous advantages via its $actor - critic$ framework to manage and optimize an online energy system.A multi-agent DRL-based approach was developed by [29] for real-time route selection and dispatching of multiple coordinated MESSs, and it had the ability to handle a hybrid continuous-discrete action space and to strengthen resiliency after extreme events. In [30], the researchers proposed a DRL-based method that coordinated the scheduling of the

MESSs and the resource allocation of micro-grids in order to minimize the overall cost of the system. In the future, a significant amount of data is expected to be generated in the fields of energy and transportation [19], which could provide significant data support for solving MESS's challenges through the application of DRL-based methods.

We have summarized the commonly used forms of energy storage, including their optimization objectives, research methods, and integration with the power grid, as shown in Table 1.

Table 1. Summary of energy management problem and power grid.

	Authors	**[31]**	**[9]**	**[10]**	**[11]**	**[32]**	**[30]**	**Ours**
	Year	**2021**	**2020**	**2021**	**2021**	**2022**	**2020**	**2023**
Modality	SESSs	✓						
	MESSs		✓	✓	✓	✓	✓	✓
Objective function	Cost/benefit	✓	✓		✓	✓		✓
	Grid peak			✓				
	Resilience						✓	
Method	Math programming		✓		✓			
	Heuristic					✓		
	Control	✓		✓				
	Deep reinforcement learning						✓	✓
Coordinated with	Power grid	✓	✓	✓		✓	✓	✓
	Renewable energy							✓

Therefore, based on the aforementioned research, most of the methods adopted to solve joint optimizations in power systems, MESSs, and other renewable energy components have been commercial solvers, heuristics, and meta-heuristics [12,20]. More recently, DRL-based methods have shown strong abilities in resolving combinatorial optimization challenges,and all have achieved remarkable results. To the best of our knowledge, however, few studies have focused on applying DRL to solve the cooperative scheduling challenges of batteries and power systems. Therefore, we proposed a joint mixed-integer programming (MIP) model for MESSs and the power grid. To address the nonlinearity and the elimination of integer variables, a hybrid strategy was proposed. A MESS model, which included 0–1 integer variables, was resolved using a DRL-based approach. The output from the DRL-based method was used as a portion of the input for the power-grid model, which was a simple linear programming (LP) model and easy to solve. The MESSs were formulated as constrained Markov decision processes (CMDPs), which served as the foundation for the DRL design. To enable the agent to make simultaneous decisions about the destination and power, a hybrid discrete-continuous action space was designed. The effectiveness of the proposed method was demonstrated through numerical experiments. Our contributions were summarized in three key areas:

(1) We developed an MIP model for the joint system of the MESSs and power grids to minimize the total operating costs. The model included constraints on the output of thermal power and renewable energy sources, the energy transmission of the power grid and the MESSs, the locations of the MESSs, etc.;
(2) We presented a formulation of the MESS challenge as a CMDP and introduce a DRL-based algorithm to make decisions in a hybrid action space that combined both discrete and continuous variables;
(3) We proposed a new linear programming (LP) model that eliminated the constraints on the MESSs in the original model by using the DRL-based method, thereby eliminating the integer-variable constraints and significantly improving the solving time. There-

fore, we proposed an algorithmic framework that combined DRL-based methods with the solutions of new LP models.

The remainder of this paper is structured as follows. In Section 2, the scheduling optimization challenge of the MESSs and the power grid is formulated as a MIP model. Section 3 describes the proposed DRL-based method. In Section 4, simulation studies are performed to validate the efficacy of the proposed approach. Section 5 presents the conclusions.

2. Mobile Energy Storage Systems and Power Grid Model

The joint optimization model for the MESSs and power grids aimed to minimize costs while considering constraints on energy transportation and operation. The objective function was comprised of two components: thermal power cost and MESS transportation cost, as follows:

$$\min \sum_{g\in G}\sum_{u\in U_g}\sum_{t\in T} C^g_{u,t}P^g_{u,t} + c^{tra}\Delta t \sum_{g\in G}\sum_{f\in G}\sum_{t\in T}\gamma_{g,f,t}, \tag{1}$$

where G denotes the electricity consumption node set and g denotes a single node; U_g denotes the thermal power unit set in node g; T represents the decision cycle, and t represents time index, where time slot length δt is set to 1 hour. The variables $C^g_{u,t}$ and $P^g_{u,t}$ indicate the generation cost of thermal power units and output, respectively. In addition, c^{tra} denotes the transportation cost per unit of time, which was set to \$20/h in this study. The expression $\gamma_{g,f,t} \in 0,1$ represents whether the MESSs were traveling between nodes g and f (1 indicated traveling and 0 indicated not traveling).

The constraints of the joint model were the following:

$$P^U_{u,\min} \le P^U_{u,t} \le P^U_{u,\max}, \quad \forall u \in U_g, g \in G, t \in T, \tag{2}$$
$$P^U_{u,t} - P^U_{u,t-1} \le RU_{u,t}, \quad \forall u \in U_g, g \in G, t \in T, \tag{3}$$
$$P^U_{u,t} - P^U_{u,t-1} \ge -RD_{u,t}, \quad \forall u \in U_g, g \in G, t \in T, \tag{4}$$
$$0 \le P^R_{r,t} \le P^R_{r,t,\max}, \quad \forall r \in R_g, g \in G, t \in T, \tag{5}$$
$$-P^{TL}_{l,t,\max} \le P^{TL}_{l,t} \le P^{TL}_{l,t,\max}, \quad \forall l \in TL, g \in G, t \in T, \tag{6}$$
$$\sum_{l\in TL} P^{TL}_{l,t} = 0, \quad \forall t \in T, \tag{7}$$
$$\sum_{u\in U_g} P^g_{u,t} + \sum_{r\in R_g} P^R_{r,t} + \sum_{l\in TL} P^{TL}_{l,t} = P^L_{g,t} + \omega_{g,t}P^{MESS}_{g,t}, \quad \forall u \in U_g, r \in R_g, g \in G, t \in T, \tag{8}$$
$$0 \le P^{MESS}_{g,t} \le \omega_{g,t}P_{max}, \quad \forall g \in G, t \in T, \tag{9}$$
$$\sum_{g\in G} \omega_{g,t} \le 1 - \sum_{f\in G} \gamma_{g,f,t}, \quad \forall g \in G, t \in T, \tag{10}$$
$$\alpha_{g,t} - \beta_{g,t} = \omega_{g,t} - \omega_{t,(t-1)}, \quad \forall g \in G, t \in T, \tag{11}$$
$$\sum_{g\in G} (\alpha_{g,t} + \beta_{g,t}) \le 1, \quad \forall t \in T, \tag{12}$$
$$\sum_{f\in G} \gamma_{g,f,t} \ge \beta_{g,t}, \quad \forall f \in G, t \in T, \tag{13}$$
$$\alpha_{f,t} - \theta_{f,t} = \sum_{g\in G} \left(\gamma_{g,f,t-1} - \gamma_{g,f,t}\right), \quad \forall f \in G, t \in T, \tag{14}$$
$$\sum_{g\in G} (\alpha_{g,t} + \theta_{g,t}) \le 1, \quad \forall t \in T, \tag{15}$$
$$\gamma_{g,f,t} \ge \gamma_{g,f,t-1} - \gamma_{g,f\left(t-H_{g,f,t}\right)}, \quad \forall g \in G, f \in G, t \in T. \tag{16}$$

where $P^U_{u,\min}$ and $P^U_{u,\max}$ denote the minimum output and maximum output of thermal power units, respectively. The variables $RU_{u,t}$ and $RD_{u,t}$ denote the increased or decreased output of the thermal units, respectively, (with ramp constraints). In addition, R_g is the set

of renewable energy resources including WP, PV, and hydropower. The actual output of the renewable energy resources and the maximum output of the renewable energy resources are expressed as $P^R_{r,t}$ and $P^R_{r,t,max}$, respectively. Furthermore, $P^{TL}_{l,t}$ and $P^{TL}_{l,t,\max}$ represent the actual transmission power and maximum transmission power of tie-line l, respectively. The expression $l = (g, f) \in TL$ denotes the tie-line between nodes g and f. The load in node g is represented as $P^L_{g,t}$.

The constraints were divided into two categories. The first category included the constraints on the power grid, including inequalities, as expressed in (2)–(8). Constraint (2) included the upper and lower limits of the thermal unit output. Constraints (3) and (4) represented the ramp constraints for thermal power units. The output of renewable energy resources was constrained by inequality, as shown in (5). The constraints on the transmission power of the tie-lines were expressed in (6) and (7). Constraint (8) represented the balance of power output for the whole joint system of the MESSs and power grids.

The second category included constraints (9)–(16) of the MESSs, with respect to [11]. Constraint (9) denoted the limit of the power output of the MESSs. The storage could not stop at a node if it was traveling between nodes and could only appear at one node at one time, as expressed in constraint (10). Equations (11)–(15) modeled the transportation status between nodes g and f, where $\alpha_{g,t} \in \{0, 1\}$ and $\beta_{g,t} \in \{0, 1\}$ were both binary variables that denoted whether the MESSs were moving to node g at time t or whether the MESSs were moving from node g at time t, respectively.In addition, $\theta_{g,t} \in \{0, 1\}$ represented an auxiliary binary variable. Constraint (11) connected the change of the locator indicators $\omega_{g,t}$ with the arrival indicators $\alpha_{g,t}$ and depart indicators $\beta_{g,t}$. The arrival and departure could not occur simultaneously, as expressed in constraint (12). Constraint (13) managed the storage depart node. Constraints (14) and (15) ensure that the arrival indicators $\alpha_{g,t}$ are equal to 1 for the arrival times; otherwise, they were equal to 0. Constraint (16) calculated the transportation time $H_{g,f,t}$ from node g to node f. In this study, the traveling time was estimated as a square matrix for all nodes.

Model (1) contained a large number of integer variables, which are known to be computationally demanding. In this proposal, we considered a machine-learning approach to solve for the integer variables. This study formulated the storage movement process as a Markov decision process (MDP) and applied a DRL-based method to solve it.

3. Deep Reinforcement Learning-Based Method

This section begins with a concise overview of the evolution of DRL-based techniques. The MESSs challenge is then framed as an MDP. The agent in the DRL-based approach was represented by an electric truck and its operator. Next, we demonstrated the integration of the DRL-based technique into the MIP model. Finally, the formulation of the action space, state space, and reward function for the DRL-based method was established. The proximal gradient projection algorithm was employed to ensure the safe exploration by the agent.

3.1. Background of DRL

The basement of the reinforcement-learning phase was the Bellman equation [33]:

$$Q^*(s, a) = \mathbb{E}_{n' \sim \varepsilon}\left[r + \gamma \max_{a'} Q^*\left(s', a'\right) \mid s, a\right], \tag{17}$$

where the optimal value of the state sequence s, represented by $Q(s, a)$, is determined by selecting the action a that maximizes the expected value of $Q(s, a)$.

Traditional reinforcement learning encounters the challenge of dimensionality. The authors of [34] addressed this issue by utilizing neural networks (NNs) to approximate the Q-value table. The Q-value was approximated using NNs, expressed as $Q(s, a; w)$, which was a close estimation of the true Q-value $Q(s, a)$. The weights of the NNs were represented

by the variable w. The weight parameters were updated through the loss function [35]:

$$L_t(w) = \left\{ Q(s_t, a_t; w) - \left[r_t + \gamma \max_{a' \in \mathcal{A}} Q\left(s_{t+1}, a'; w_t\right) \right] \right\}^2. \tag{18}$$

The maximization calculation in Equation (18) is computationally inefficient for continuous action spaces due to the non-convex nature of the function $Q(s, a; w)$ with respect to action a. This renders the maximization calculation NP-hard in the worst-case scenario [36]. To overcome this challenge, ref. [37] introduced the deterministic policy gradient (DPG) theorem which utilizes policy-based methods for continuous action spaces through the implementation of deterministic policies $\mu_\theta : S \to \mathcal{A}$. The objective of policy gradient methods is to discover a policy π_θ that optimizes the expected reward. The objective of the policy, $J(\pi_\theta)$, is then updated through gradient descent [38]:

$$\nabla_\theta J(\mu_\theta) = \mathop{\mathbb{E}}_{s \sim \rho^{\mu_\theta}} \left[\nabla_\theta \mu_\theta(s) \nabla_a Q^{\mu_\theta}(s, a)|_{a=\mu_\theta(s)} \right]. \tag{19}$$

3.2. Algorithm Framework

The *actor* − *critic* framework has proven to be effective in various domains, such as energy storage systems [26] and the game Go [39]. We adopted a similar structure in this proposal. The *Actor* network was a strategic network designed to determine the power parameters of the MESSs, which had a continuous feasible range. The *Critic* network assessed the actor–network's parameter choices and decided on other discrete variables, such as destination selection, charging and discharging selection, etc.

The hybrid action space $\mathcal{A}$ was defined according to [36], as follows:

$$\mathcal{A} = \{(k, x_k) \mid x_k \in \mathcal{X}_k \text{ for all } k \in K\}, \tag{20}$$

which consists of discrete action $k \in \mathcal{K}$ and continuous action $x_k \in \mathcal{X}_k$. The Bellman equation transforms into [36], as follows:

$$Q\left(s_t, k_t, x_{k_t}\right) = \mathop{\mathbb{E}}_{r_t, s_{t+1}} \left[r_t + \gamma \max_{k \in K} Q\left(s_{t+1}, k, x_k^Q(s_{t+1})\right) \mid s_t = s \right]. \tag{21}$$

The value-based network's weights are denoted as ω, and the policy-based network's weights are denoted as θ. In each step t of the updating process, ω_t and θ_t were updated, respectively. First, the weight parameter, ω, was estimated by employing the gradient descent to minimize the mean-squared Bellman error, which was similar to the DQN method. Then, ω and θ were fixed by maximizing $Q(s, k, x_k(s; \theta); \omega)$. The loss functions were formulated as follows [40]:

$$\ell_t^Q(\omega) = \frac{1}{2} \left[Q\left(s_t, k_t, x_{k_t}; \omega\right) - y_t \right]^2, \tag{22}$$

$$\ell_t^\Theta(\theta) = -\sum_{k=1}^{K} Q(s_t, k, x_k(s_t; \theta); \omega_t), \tag{23}$$

where $y_t = r_t + \gamma \max_{k' \in K} Q(s_{t'}, k, x_k(s_{t'}, \theta_t); \omega_t)$.

According to [41], there were three approaches for utilizing machine learning in combinatorial optimization challenges: using machine learning on its own, integrating machine learning with traditional optimization algorithms, and iteratively combining optimization and machine learning. For this proposal, we employed the DRL-based approach to efficiently eliminate integer variables in model (1). The objective function of model 1 served as the reward for reinforcement learning, and the DRL-based method was used to obtain the integer variables in constraints (9)–(16). The value of $\omega_{g,t}$ in constraint (8) was passed as an input to the new model.

The new model was described by the following:

$$\begin{aligned} &\min \sum_{g\in G}\sum_{u\in U_g}\sum_{t\in T} C^g_{u,t}P^g_{u,t} + C^{tra} \\ &s.t. \quad \text{constraints (2)} - \text{(8) in model (1)}, \end{aligned} \tag{24}$$

where C^{tra} is the total transportation cost of the MESSs.

The algorithm framework is depicted in Figure 1. State S_t (P1) consisted of the output of thermal units and renewable energy at grid nodes, while State S_t (P2) included the remaining energy of the MESSs, the cost of thermal power units, and the location information. State S_t was the combination of these five elements.

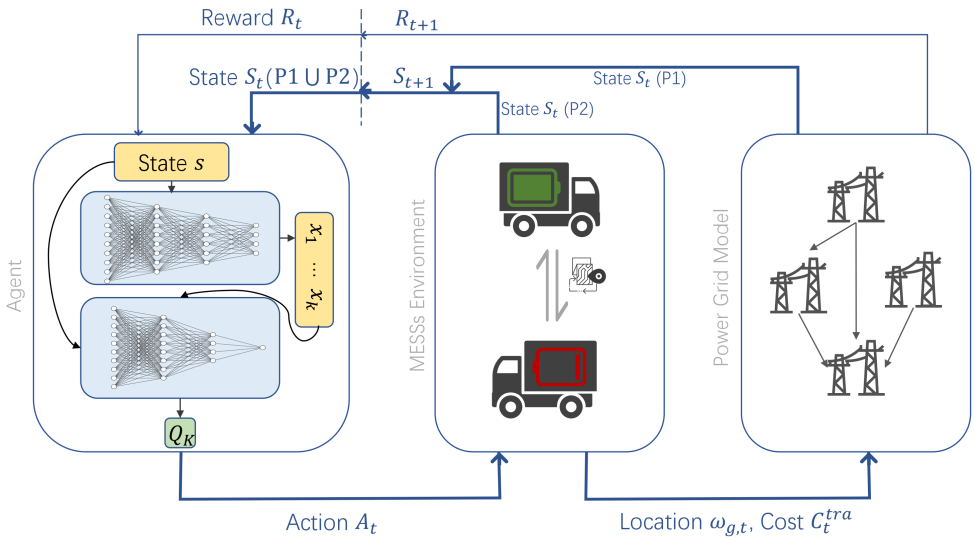


Figure 1. Framework of optimization method for the joint model of MESSs and power grids.

3.3. Constrained MDP Formulation

According to model (1), this study modeled the MESSs in this model as MDPs. However, due to the capacity constraint of the battery packs in the MESSs, the processes became CMDPs. To address this, we designed a hybrid action space and a state space for the CMDPs. A proximal gradient projection algorithm was proposed to ensure safe exploration by the agent.

3.3.1. Action Space $\mathcal{A}$, State Space $\mathcal{S}$, and Reward Function $\mathcal{R}$

Action space $\mathcal{A}$ was composed of two levels: The discrete action space that included a chosen charging/discharging/holding action and the chosen destination node; and the continuous action that included the output power parameters of the MESSs. At time t, the action a_t was represented by $\left((g_t, cdh_t)|P^{cdh}_{g,t}\right)$, where g is the chosen destination node, cdh represents the charging/discharging/holding choice, and $P^{cdh}_{g,t}$ is the corresponding power at node g at time t. The range of actions and their values were defined by the following:

$$\mathcal{A} = \begin{cases} g \in \{g_1, g_2, \ldots, g_N\} \\ cdh \in \{-1, 0, 1\} \\ P = \begin{cases} (0, P_{max}], cdh = -1 \text{ or } chd = 1 \\ 0, cdh = 0 \end{cases} \end{cases}, \tag{25}$$

where N is the node's number, and $cdh = -1$, 0, and 1 means discharging, holding, and charging, respectively.

The state space $\mathcal{S}$ consisted of two parts, as illustrated in Figure 1. The variable $\mathcal{S}_t(P1)$ described the state of the power grid, and $\mathcal{S}_t(P2)$ described the state related to the MESS. At time t, $\mathcal{S}$ could be defined as $\mathcal{S}_t = \left(P_{u,t}, RE_{g,t}, E_{g,t}, Load_{g,t}, g\right)$, where $RE_{g,t}$ is the renewable energy resources at node g at time t, $E_{g,t}$ is the remaining energy in the MESSs at node g at time t, and $Load_{g,t}$ represents the load at node g. There was a capacity constraint for the MESSs, $\underline{E} \leq E_{g,t} \leq \bar{E}$, which indicated the battery pack had an upper and lower limit on its capacity, represented by $\underline{E}$ and $\bar{E}$, respectively.

The reward $\mathcal{R}$ was based on the objective value of the linear programming (LP) model (24). The reward in RL was a cumulative value, as depicted in Equation (21), where $Q(s,a) = \sum_{t \in T} \gamma^t r_t$. In this study, the objective value of model (24) was used as the reward: $r_t = -Obj_{m2}$, where Obj_{m2} represents the optimal objective value of model (24).

Theorem 1. *In a finite MDP, if $\sum_{t \in T} -Obj_{m2,t}$ was the maximum when training converged with a policy, then each $-Obj_{m2,t}$ $t \in T$ was the maximum in the policy.*

Proof of Theorem 1. If ϕ is a policy, then

$$q_\pi(s^*, a^*) \geq v_\pi(s^*).$$

Suppose, then, that there is a policy $\phi^{'}$:

$$v_{\pi^{'}}(s) \geq v_\pi(s) \text{ for all } s,$$
$$\pi^{'}(a \mid s) = 1(s^* = s)1(a = a^*) + 1(s^* \neq s)\pi(a \mid s).$$

We considered $v^*(s) = \max_{\pi \in \Pi} v_\pi(s)$ and $q^*(s,a) = \max_{\pi \in \Pi} q_\pi(s,a)$. Now extrapolating backward, there existed s^*, resulting in the following:

$$\max_a q^*(s^*, a) > v^*(s^*).$$

Suppose there existed a policy $\pi^{''}$, as well as a^*:

$$q_{\pi^{''}}(s^*, a^*) = \max_a q^*(s^*, a)$$

while v^* was the maximum over all π, for $\pi = \pi^{''}$, we obtained the following inequality:

$$q_{\pi^{''}}(s^*, a^*) > v_{\pi^{'}}(s^*).$$

Therefore, we obtained the following:

$$v_{\pi^{'}}(s^*) \geq q_{\pi^{''}}(s^*, a^*) > v_{\pi^{'}}(s^*).$$

However, $q_{\pi^{''}}(s^*, a^*)$ was strictly larger than $v_\pi(s^*)$ for any policy π, so we obtained the following contradiction:

$$v_{\pi^{'}}(s^*) > v_{\pi^{'}}(s^*)$$

Therefore, if a policy was optimal, then it would be optimal for all states in all steps. □

3.3.2. State-Updating Process

Based on the design of $\mathcal{S}$ and $\mathcal{A}$, new states could be observed after taking actions via $\mathcal{A}$ in both the MESS environment and the power-grid environment. The state-updating equations for time-step t were described separately for the three cases of charging, discharging, and holding, as follows:

$$\text{charge}: \begin{cases} P_{u,t^{'}} = P^{LP}_{u,t} \\ RE_{g,t^{'}} = \mathcal{M}_{re}(RE_{g,t}) + E_{g,t} \\ E_{g,t^{'}} = E_{g,t} + P_{g,t}\Delta h \\ Load_{g,t^{'}} = \mathcal{M}_{load}(Load_{g,t}) \\ g^{'} = Action(g) \end{cases},$$

$$\text{discharge}: \begin{cases} P_{u,t'} = P_{u,t}^{LP} \\ RE_{g,t'} = \mathcal{M}_{re}(RE_{g,t}) - E_{g,t} \\ E_{g,t'} = E_{g,t} - P_{g,t}\Delta h \\ Load_{g,t'} = \mathcal{M}_{load}(Load_{g,t}) \\ g^{'} = Action(g) \end{cases},$$

$$\text{hold}: \begin{cases} P_{u,t'} = P_{u,t}^{LP} \\ RE_{g,t'} = \mathcal{M}_{re}(RE_{g,t}) \\ E_{g,t'} = E_{g,t} \\ Load_{g,t'} = \mathcal{M}_{load}(Load_{g,t}) \\ g^{'} = Action(g) \end{cases},$$

where $\mathcal{M}_{re}$ and $\mathcal{M}_{load}$ are the matrices that hold the renewable energy sources and loads, respectively.

3.3.3. Proximal-Gradient Algorithm

In accordance with our previous discussion in Section 3.3.1, we studied the MESSs challenge under the CMDP framework. Several approaches were available for ensuring safe exploration by the agent, including the primal-dual algorithm, the adaptive-penalty method, and the gradient-projection algorithm. However, for the purposes of this study, we deemed the proximal-gradient algorithm to be both readily implementable and computationally efficient.

In the normal gradient-descent process, the iterative equation was the following:

$$x_{t+1} = x_t - \alpha_t \cdot f_t(x_t), \tag{26}$$

where α_t is the step length in step t. When the solution $\widetilde{x}_t$ exceeded its feasible domain, we needed to project it back into the feasible domain. Therefore, the iterative equation was transformed into the following:

$$x_t = \mathcal{P}_X(\widetilde{x}_t) := \arg\min\left\{ \|x - \widetilde{x}_t\|^2 : x \in X \right\}, \tag{27}$$

where $\mathcal{P}_X$ is the projection operator, which was chosen as the Euclidean distance function in our study, and $\widetilde{x}_t := x_{t-1} - \gamma_n \hat{\nabla}_x f(x_{t-1})$ denoted the original infeasible solution.

3.4. Learning Process

As mentioned in Section 3.2, this proposal employed the "*actor* − *critic*" (A-C) framework to train the policy NN and value NN, respectively. To ensure the network was convergent and stable, we used target networks that were added to each NN of the A-C structure and added an experience replay pool, such as the double-DQN(DDQN) [42]. The learning process was also similar to the update process of the DDQN.

The required inputs for the algorithm included the initialized parameters such as exploration parameter ϵ, mini-batch size B, a probability distribution ξ, etc. The capacity N of the experience replay memory $\mathcal{D}$; the parameters of the *Actor* network Θ and the *Critic* network Q; and their target networks $\hat{\Theta}$ and $\hat{Q}$, respectively, also needed to be initialized.

At the beginning of each episode $i \in I$, where I is the maximum training time, the agent acquired an initial state by observation and computed continuous action parameters through network Θ: $x_k \leftarrow x_k(s_t, \theta_t)$. Then, it selected an action $a_t = \left(k_t, x_{k_t}\right)$, according to the ϵ-greedy policy:

$$a_t = \begin{cases} \text{a sample from distribution } \xi \text{ with probability } \epsilon \\ \left(k_t, x_{k_t}\right) \text{such that } k_t = \arg\max_{k \in K} Q(s_t, k, x_k; \omega_t) \text{ with probability } 1 - \epsilon \end{cases}$$

in each decision time-step t. Then, we performed the action in the MESSs and power grid environment to define the new state s_{t+1} and the reward r_t, which was obtained by solving the model (24).

Next, the transition tuple $[s_t, a_t, r_t, s_{t+1}]$ was stored in $\mathcal{D}$. Through sampling a batch $\{s_b, a_b, r_b, s_{b+1}\}_{b\in[B]}$ of size B, the target y_b could be calculated via equation

$$y_b = \begin{cases} r_b & \text{if } s_{b+1} \text{ is the terminal state,} \\ r_b + \max_{k\in K} \gamma Q(s_{b+1}, k, x_k(s_{b+1}, \theta_t); \omega_t) & \text{if otherwise.} \end{cases}$$

Then, the stochastic gradients $\nabla_\omega \ell_t^Q(\omega)$ and $\nabla_\theta \ell_t^\Theta(\theta)$ were calculated through $\{y_b, s_b, a_b\}_{b\in[B]}$, as shown in Equations (22) and (23). The weights were updated by $\theta_{t+1} \leftarrow \theta_t - \beta_t \nabla_\theta \ell_t^\Theta(\theta_t)$. In the above processes, the proximal gradient projection mentioned in Section 3.3.3 was applied. We used a soft update (Polyak averaging) method to update the target networks $\hat{Q} = (1-\tau)Q + \tau\hat{Q}$ and $\hat{\Theta} = (1-\tau)\Theta + \tau\hat{\Theta}$, where $\tau \in [0,1]$ is a hyper-parameter, usually $\tau \ll 1$.

4. Case Studies

In order to verify the effectiveness of the proposed strategy, case studies were implemented for the integrated model, utilizing simulation data. Our comparison of the energy output after incorporating MESSs demonstrated that the proposed strategy was effective in reducing the proportion of thermal power output and enhancing the utilization of renewable energy sources.

4.1. Experiment Settings

The training process for the 10-node scenario was approximately 5 h in length, on a desktop computer with an NVIDIA GTX 3080 GPU and an Intel i7-13700KF CPU, and approximately 6 h on a server with an NVIDIA Tesla P100.The solver for the MIP and LP models was Gurobi Optimizer, version 9.11.

The MESSs comprised an electric truck, a battery pack, and an operator. The unit movement cost c^{TRA} incorporated all expenses associated with the movement, including the charging costs of the electric truck and the operator's wages. The experimental arithmetic was represented by a topological diagram comprising 10 nodes that were interconnected by electrical conduits. The hyper-parameters of the experiment were subject to multiple simulations to find a balance between convergence speed and stability. The main simulation parameters are shown in Table 2.

Table 2. Experiment Parameters

Parameter Names	Value
Maximum capacity $\bar{E}$ (MWh)	27
Maximum charge/discharge power pwr_{MAX} (MW)	27
Maximum number of nodes n	10
Maximum number of thermal power units	3
Charging and discharging efficiency η	0.95
Transportation cost per unit of time c^{TRA} (\$/h)	20
Random exploration parameter ε	0.9–0.1
Batch size B	128
Discount factor γ	0.95
Probability distribution in random exploration ξ	U(0,1)
Soft updating parameter τ	0.1
Sizes of three layers of neural networks	[256,128,64]
Learning rate in policy network lr_p	1×10^{-6}
Learning rate in value network lr_v	1×10^{-4}

4.2. Results Analysis

Figure 2 depicts the convergence of the cumulative rewards achieved by the proposed DRL-based algorithm during the training phase. The x-axis represented the number of iteration rounds, with each increment on the axis representing 100 iterations. The y-axis displayed the accumulated cost, which was presented as a negative value. The graph demonstrated that the training reached convergence after approximately 1500 iterations. The training process was approximately 5 h for 7000 iterations, whereas the execution time of the trained network was virtually insignificant. By utilizing a trained network, we observed a significant improvement in computational efficiency.

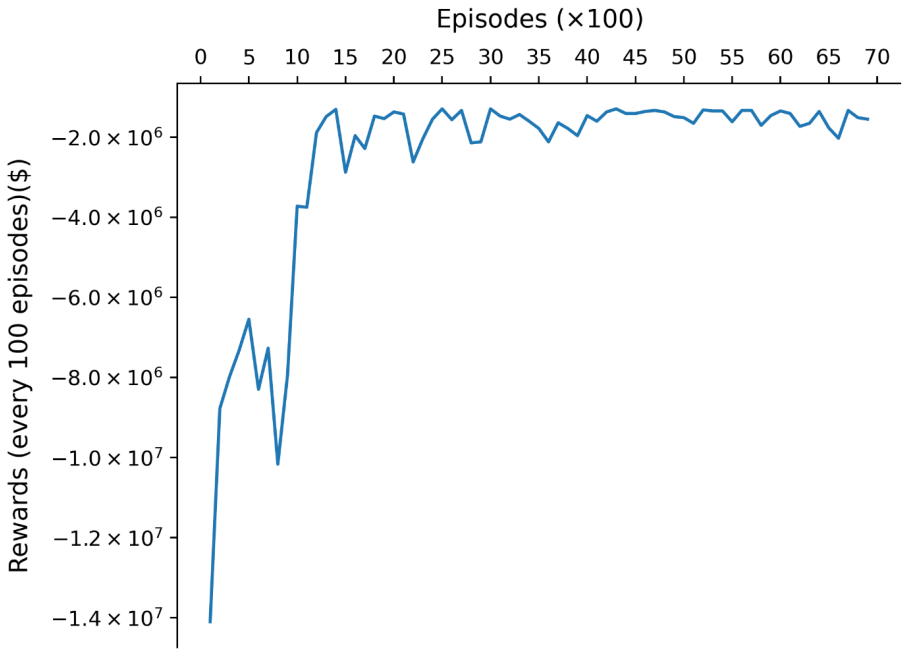
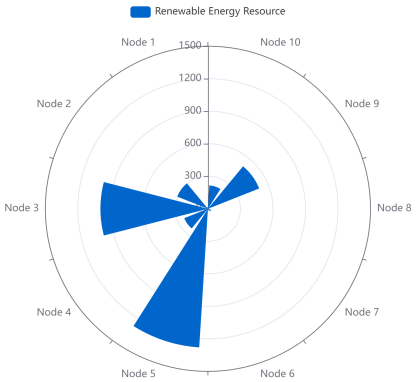


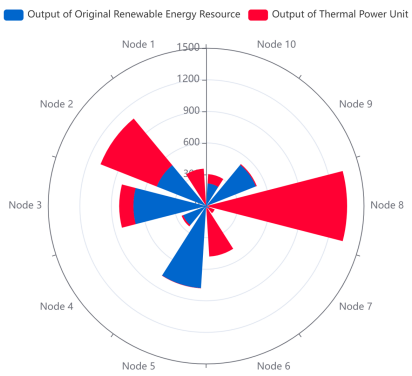
Figure 2. Episodic average reward in every 100 episodes in a scenario with 10 nodes.

Figure 3 shows the alteration in the distribution of renewable energy and thermal unit outputs within the system. As depicted in Figure 3a, certain nodes, such as nodes 3 and 8, exhibited a richer availability of renewable energy sources, whereas others, such as nodes 1, 6, and 7, exhibited virtually no presence of such resources. To cater to the electricity demand of these nodes with limited renewable energy, it was necessary to utilize thermal power units. Figure 3b presents a polar plot, demonstrating the distribution of renewable energy and thermal unit outputs in the absence of a connection to the MESS system. The figure clearly illustrates that some nodes, such as node 5, exhibited a predominant presence of renewable energy output, while others, such as node 8, were dominated by thermal power generation. Overall, the system exhibited a relatively substantial proportion of thermal power generation.

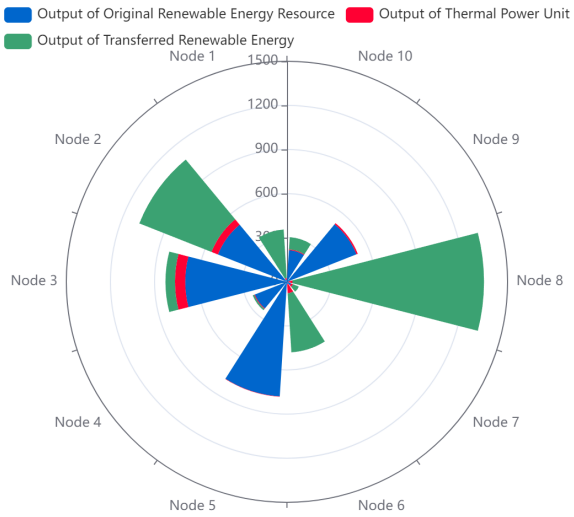
Figure 3c presents a polar plot that displays the optimal energy distribution among various nodes. The blue, red, and green segments in the plot depicted the local renewable energy sources, the energy generated by thermal units, and the energy transmitted through the MESSs, respectively. As shown in the plot, certain nodes, such as nodes 5 and 9, possessed abundant local renewable energy resources and, thus, exhibited a higher proportion of local energy output. These nodes were also the main charging nodes and were not typically involved in discharging the MESSs.Comparatively, nodes such as node 2 had both local renewable energy resources and energy transmitted by the MESSs contributing to their power supply. Furthermore, nodes such as node 8, which primarily relied on the energy transmitted by the MESSs, effectively reduced their share of the thermal power-unit output.



(**a**) Renewable Energy Resource at Each Node in a 10-Node Scenario



(**b**) Share of Energy Output of the Original System at Each Node in a 10-Node Scenario



(**c**) Share of Energy Output of the System after Integration of the MESSs at Each Node in a 10-Node Scenario

Figure 3. The Output of Thermal Power Units, Original Renewable Energy, and Transferred Renewable Energy at Each Node in a 10-Node Scenario.

Figure 4 presents a chord diagram that displays the cumulative energy transfer within the decision cycle of the MESSs. The diagram provided a visual representation of the flow of energy transfer and its accumulation during the decision cycle. The direction of the arrows represents the flow of energy transfer. As depicted in the figure, some nodes were designated as primary charging nodes, such as nodes 3 and 5, while others served as primary discharging nodes, such as nodes 1 and 8. This was because these nodes, such as 3 and 5, possessed abundant renewable energy sources, which was a result of their geographical locations and other contributing factors. However, when these nodes were not connected to the MESSs, the surplus of renewable energy resources could lead to issues such as excess light or wind. By incorporating these nodes into the MESSs,the proportion of thermal power output was significantly reduced, thus improving the utilization of renewable energy sources. Some nodes, such as nodes 3 and 8, served as both supply and demand nodes, which was due to the intermittent nature of renewable energy and the mismatch between the supply of renewable energy and the peak demand for electricity. For example, the power generation capacity of PV was related to factors such as sunlight duration and weather and reached its peak output during sunny afternoons, but the local demand may not be able to consume all the generated electricity during this time period.

Therefore, the excess electricity would then be transported to nearby nodes to meet the electricity demand of other nodes. In the evening, when the power output of PV decreased, a hydroelectric power supply from neighboring nodes could be required.

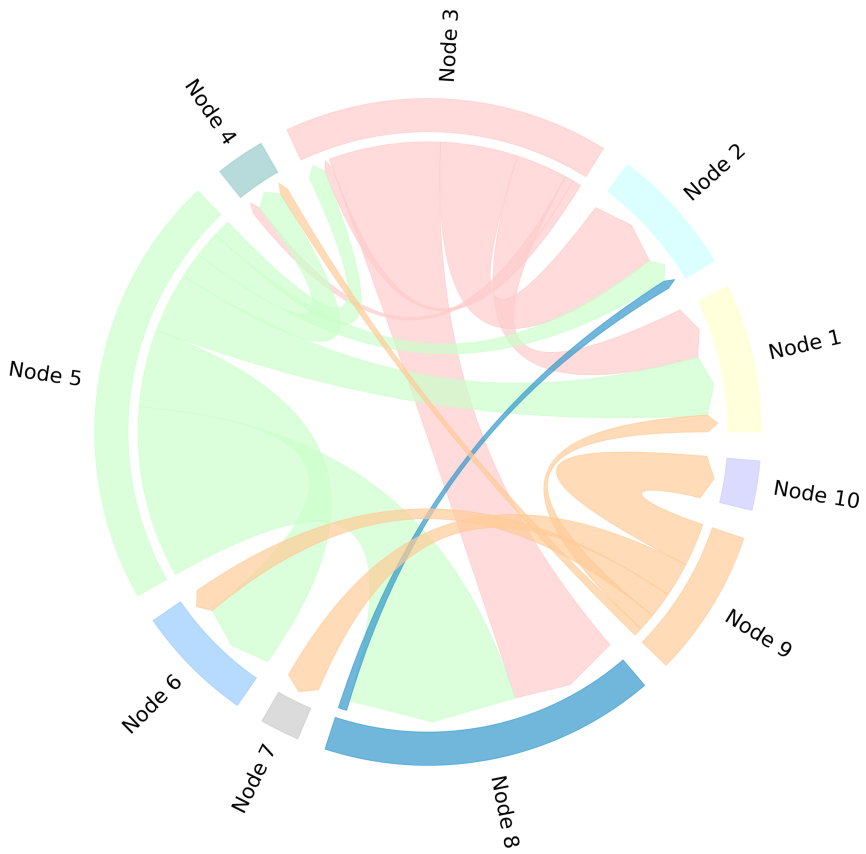


Figure 4. Optimal cumulative total renewable energy transfer in a 10-node scenario.

By conducting numerous experiments, we demonstrated that by integrating the MESSs into the power grid, it was feasible to transfer surplus renewable energy from nodes with an abundance of such resources, to nodes with limited resources. This resulted in a reduction in thermal power generation in the latter, thereby validating the proposed model and algorithm.

Furthermore, our proposed mobile energy storage system-grid joint optimization model has the potential to address network contingency events. First, our study on coupling mobile energy storage systems with the power grid contributes to achieving higher renewable energy penetration levels [43,44]. In the event of network contingencies (e.g., equipment failure, fluctuations in power demand, etc.), mobile energy storage systems can quickly adjust their output to balance supply and demand imbalances in the grid. Moreover, mobile energy storage systems can transfer electrical energy from one location to another within the grid in a short period of time, alleviating local transmission congestion issues and improving the stability of the entire grid. Second, by optimizing the dispatch strategy of mobile energy storage systems, we can allocate renewable energy resources reasonably across different time periods to meet grid demand. This will help mitigate power supply shortages or overloads caused by contingency events, further enhancing the stability and reliability of the grid [29]. Lastly, our research also indicates that mobile energy storage systems can help reduce the operational costs of the entire power system [45]. By utilizing renewable energy resources more efficiently, we can decrease our reliance on conventional fossil fuel-based generation, thus lowering the cost losses associated with network contingencies.

5. Conclusions

In this study, we developed a mixed-integer nonlinear programming (MINP) model that coupled MESSs with a power grid to balance a region with an uneven distribution of renewable energy. We modeled the MESSs as constrained Markov decision processes (CMDPs) and proposed a framework based on a deep reinforcement learning (DRL) algorithm that considered the discrete-continuous hybrid action space of the MESSs. We solved the constraint on battery capacity in the CMDP by applying a proximal-gradient-projection algorithm. Based on this, we reformulated the original MINP model as a linear programming (LP) model to arrive at our solution framework. The case study showed that our proposed algorithm and framework effectively improved the utilization of renewable energy and reduced the generation costs of thermal power units.

The DRL-based algorithm we proposed has good scalability and application prospects in energy challenges with discrete-continuous hybrid decision-action spaces. Furthermore, our approach of using a DRL-based method to eliminate integer decision variables and nonlinear constraints provided insight for solving the MINP challenges. However, to highlight the main innovation points of this paper, we simplified some constraints, such as overlooking power-flow constraints and the aging characteristics of batteries. In future research, we will consider these additional constraints and objective functions more carefully, including the initial investment cost for implementing MESSs, and attempt to optimize scheduling decisions over longer time periods.

Author Contributions: Conceptualization, Y.D. and J.W.; methodology, Y.D. and J.W.; software, Y.D.; validation, Y.D.; formal analysis, Y.D., X.C. and J.W.; investigation, Y.D. and X.C.; resources, Y.D.; data curation, X.C.; writing—original draft preparation, Y.D.; writing—review and editing, Y.D., X.C. and J.W.; visualization, Y.D.; supervision, J.W.; project administration, Y.D. and J.W.; funding acquisition, J.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported, in part, by the National Key Research and Development Program, under Grant 2022YFB2405600; and, in part, by the National Natural Science Foundation of China, under Grants 72131001, T2121002, and 52277092.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: This work was supported by the High-performance Computing Platform of Peking University.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MESSs	Mobile Energy Storage Systems
WP	Wind Power
PV	Photovoltaic
MIP	Mixed-Integer Programming
LP	Linear Programming
DRL	Deep Reinforcement Learning
ESSs	Energy Storage Systems
SESS	Stationary Energy Storage System
EV	Electric Vehicle
MILP	Mixed-Integer Linear Programming
MDP	Markov Decision Process
CMDP	Constrained Markov Decision Process
A-C	Actor-Critic
NNs	Neural Networks

References

1. Olabi, A.; Abdelkareem, M.A. Renewable energy and climate change. *Renew. Sustain. Energy Rev.* **2022**, *158*, 112111.
2. Elavarasan, R.M.; Shafiullah, G.; Padmanaban, S.; Kumar, N.M.; Annam, A.; Vetrichelvan, A.M.; Mihet-Popa, L.; Holm-Nielsen, J.B. A comprehensive review on renewable energy development, challenges, and policies of leading Indian states with an international perspective. *IEEE Access* **2020**, *8*, 74432–74457.
3. Qazi, A.; Hussain, F.; Rahim, N.A.; Hardaker, G.; Alghazzawi, D.; Shaban, K.; Haruna, K. Towards sustainable energy: A systematic review of renewable energy sources, technologies, and public opinions. *IEEE Access* **2019**, *7*, 63837–63851.
4. Cole, W.J.; Greer, D.; Denholm, P.; Frazier, A.W.; Machen, S.; Mai, T.; Vincent, N.; Baldwin, S.F. Quantifying the challenge of reaching a 100% renewable energy power system for the United States. *Joule* **2021**, *5*, 1732–1748.
5. Impram, S.; Nese, S.V.; Oral, B. Challenges of renewable energy penetration on power system flexibility: A survey. *Energy Strategy Rev.* **2020**, *31*, 100539.
6. Kebede, A.A.; Kalogiannis, T.; Van Mierlo, J.; Berecibar, M. A comprehensive review of stationary energy storage devices for large scale renewable energy sources grid integration. *Renew. Sustain. Energy Rev.* **2022**, *159*, 112213.
7. Park, S.; Ahn, J.; Kang, T.; Park, S.; Kim, Y.; Cho, I.; Kim, J. Review of state-of-the-art battery state estimation technologies for battery management systems of stationary energy storage systems. *J. Power Electron.* **2020**, *20*, 1526–1540.
8. Maestre, V.; Ortiz, A.; Ortiz, I. Challenges and prospects of renewable hydrogen-based strategies for full decarbonization of stationary power applications. *Renew. Sustain. Energy Rev.* **2021**, *152*, 111628.
9. Saboori, H.; Jadid, S. Optimal scheduling of mobile utility-scale battery energy storage systems in electric power distribution networks. *J. Energy Storage* **2020**, *31*, 101615.
10. Kucevic, D.; Englberger, S.; Sharma, A.; Trivedi, A.; Tepe, B.; Schachler, B.; Hesse, H.; Srinivasan, D.; Jossen, A. Reducing grid peak load through the coordinated control of battery energy storage systems located at electric vehicle charging parks. *Appl. Energy* **2021**, *295*, 116936.
11. He, G.; Michalek, J.; Kar, S.; Chen, Q.; Zhang, D.; Whitacre, J.F. Utility-scale portable energy storage systems. *Joule* **2021**, *5*, 379–392.
12. He, G.; Chen, X.; Yang, Y.; Wang, J.; Song, J. Hybrid Portable and Stationary Energy Storage Systems with Battery Charging and Swapping Coordination. In Proceedings of the 2022 IEEE/IAS Industrial and Commercial Power System Asia (I&CPS Asia), Shanghai, China, 8–11 July 2022; pp. 1465–1470.
13. Abdeltawab, H.H.; Mohamed, Y.A.R.I. Mobile energy storage scheduling and operation in active distribution systems. *IEEE Trans. Ind. Electron.* **2017**, *64*, 6828–6840.
14. Nazemi, M.; Dehghanian, P.; Lu, X.; Chen, C. Uncertainty-aware deployment of mobile energy storage systems for distribution grid resilience. *IEEE Trans. Smart Grid* **2021**, *12*, 3200–3214.
15. Ebadi, R.; Yazdankhah, A.S.; Kazemzadeh, R.; Mohammadi-Ivatloo, B. Techno-economic evaluation of transportable battery energy storage in robust day-ahead scheduling of integrated power and railway transportation networks. *Int. J. Electr. Power Energy Syst.* **2021**, *126*, 106606.
16. Sexauer, J.M.; Mohagheghi, S. Voltage quality assessment in a distribution system with distributed generation—A probabilistic load flow approach. *IEEE Trans. Power Deliv.* **2013**, *28*, 1652–1662.
17. Eyer, J.; Corey, G. Energy storage for the electricity grid: Benefits and market potential assessment guide. *Sandia Natl. Lab.* **2010**, *20*, 5.
18. Abolhassani, M.H.; Safdarian, A. Electric Vehicles as Mobile Energy Storage Devices to Alleviate Network Congestion. In Proceedings of the 2019 Smart Grid Conference (SGC), Tehran, Iran, 18–19 December 2019; pp. 1–5.
19. Song, J.; He, G.; Wang, J.; Zhang, P. Shaping future low-carbon energy and transportation systems: Digital technologies and applications. *iEnergy* **2022**, *1*, 285–305.
20. Pulazza, G.; Zhang, N.; Kang, C.; Nucci, C.A. Transmission planning with battery-based energy storage transportation for power systems with high penetration of renewable energy. *IEEE Trans. Power Syst.* **2021**, *36*, 4928–4940.
21. Sun, Y.; Li, Z.; Shahidehpour, M.; Ai, B. Battery-based energy storage transportation for enhancing power system economics and security. *IEEE Trans. Smart Grid* **2015**, *6*, 2395–2402.
22. Kwon, S.Y.; Park, J.Y.; Kim, Y.J. Optimal V2G and route scheduling of mobile energy storage devices using a linear transit model to reduce electricity and transportation energy losses. *IEEE Trans. Ind. Appl.* **2019**, *56*, 34–47.
23. Yan, G.; Liu, D.; Li, J.; Mu, G. A cost accounting method of the Li-ion battery energy storage system for frequency regulation considering the effect of life degradation. *Prot. Control. Mod. Power Syst.* **2018**, *3*, 1–9.
24. Scarabaggio, P.; Carli, R.; Cavone, G.; Dotoli, M. Smart control strategies for primary frequency regulation through electric vehicles: A battery degradation perspective. *Energies* **2020**, *13*, 4586.
25. Li, W.; Cui, H.; Nemeth, T.; Jansen, J.; Uenluebayir, C.; Wei, Z.; Zhang, L.; Wang, Z.; Ruan, J.; Dai, H.; et al. Deep reinforcement learning-based energy management of hybrid battery systems in electric vehicles. *J. Energy Storage* **2021**, *36*, 102355.
26. Liang, Y.; Ding, Z.; Zhao, T.; Lee, W.J. Real-time operation management for battery swapping-charging system via multi-agent deep reinforcement learning. *IEEE Trans. Smart Grid* **2022**, *14*, 559–571.
27. Kofinas, P.; Dounis, A.; Vouros, G. Fuzzy Q-Learning for multi-agent decentralized energy management in microgrids. *Appl. Energy* **2018**, *219*, 53–67.

28. Hua, H.; Qin, Y.; Hao, C.; Cao, J. Optimal energy management strategies for energy Internet via deep reinforcement learning approach. *Appl. Energy* **2019**, *239*, 598–609.
29. Wang, Y.; Qiu, D.; Strbac, G. Multi-agent deep reinforcement learning for resilience-driven routing and scheduling of mobile energy storage systems. *Appl. Energy* **2022**, *310*, 118575.
30. Yao, S.; Gu, J.; Zhang, H.; Wang, P.; Liu, X.; Zhao, T. Resilient load restoration in microgrids considering mobile energy storage fleets: A deep reinforcement learning approach. In Proceedings of the 2020 IEEE Power & Energy Society General Meeting (PESGM), Montreal, QC, Canada, 2–6 August 2020; pp. 1–5.
31. Kebede, A.A.; Coosemans, T.; Messagie, M.; Jemal, T.; Behabtu, H.A.; Van Mierlo, J.; Berecibar, M. Techno-economic analysis of lithium-ion and lead-acid batteries in stationary energy storage application. *J. Energy Storage* **2021**, *40*, 102748.
32. Yu, Z.; Dou, Z.; Zhao, Y.; Xie, R.; Qiao, M.; Wang, Y.; Liu, L. Grid Scheduling Strategy Considering Electric Vehicles Participating in Multi-microgrid Interaction. *J. Electr. Eng. Technol.* **2022**, 1–16. https://doi.org/10.1007/s42835-022-01294-x.
33. Bellman, R. A Markovian decision process. *J. Math. Mech.* **1957**, *6*, 679–684.
34. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Graves, A.; Antonoglou, I.; Wierstra, D.; Riedmiller, M. Playing atari with deep reinforcement learning. *arXiv* **2013**, arXiv:1312.5602.
35. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G.; et al. Human-level control through deep reinforcement learning. *Nature* **2015**, *518*, 529–533.
36. Xiong, J.; Wang, Q.; Yang, Z.; Sun, P.; Han, L.; Zheng, Y.; Fu, H.; Zhang, T.; Liu, J.; Liu, H. Parametrized deep q-networks learning: Reinforcement learning with discrete-continuous hybrid action space. *arXiv* **2018**, arXiv:1810.06394.
37. Silver, D.; Lever, G.; Heess, N.; Degris, T.; Wierstra, D.; Riedmiller, M. Deterministic policy gradient algorithms. In Proceedings of the International Conference on Machine Learning, Beijing, China, 21–26 June 2014; pp. 387–395.
38. Sutton, R.S.; McAllester, D.; Singh, S.; Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. *Adv. Neural Inf. Process. Syst.* **1999**, *12*, 1–7.
39. Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A.; et al. Mastering the game of go without human knowledge. *Nature* **2017**, *550*, 354–359.
40. Bester, C.J.; James, S.D.; Konidaris, G.D. Multi-pass q-networks for deep reinforcement learning with parameterised action spaces. *arXiv* **2019**, arXiv:1905.04388.
41. Bengio, Y.; Lodi, A.; Prouvost, A. Machine learning for combinatorial optimization: A methodological tour d'horizon. *Eur. J. Oper. Res.* **2021**, *290*, 405–421.
42. Van Hasselt, H.; Guez, A.; Silver, D. Deep reinforcement learning with double q-learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016; Volume 30.
43. Dugan, J.; Mohagheghi, S.; Kroposki, B. Application of mobile energy storage for enhancing power grid resilience: A review. *Energies* **2021**, *14*, 6476.
44. Wang, Y.; Rousis, A.O.; Strbac, G. Resilience-driven optimal sizing and pre-positioning of mobile energy storage systems in decentralized networked microgrids. *Appl. Energy* **2022**, *305*, 117921.
45. Ahmadi, S.E.; Marzband, M.; Ikpehai, A.; Abusorrah, A. Optimal stochastic scheduling of plug-in electric vehicles as mobile energy storage systems for resilience enhancement of multi-agent multi-energy networked microgrids. *J. Energy Storage* **2022**, *55*, 105566.