processes

MDPI

Article

Deep Reinforcement Learning-Based Joint Low-Carbon Optimization for User-Side Shared Energy Storage–Distribution Networks

Lihua Zhong [1], Tong Ye [2,3], Yuyao Yang [1,*], Feng Pan [1], Lei Feng [1], Shuzhe Qi [1] and Yuping Huang [2,3]

[1] Metrology Center of Guangdong Power Grid Co., Ltd., Qingyuan 511545, China; zhonglihua0126@dg.gd.csg.cn (L.Z.); pf6601@163.com (F.P.); 18879764689@163.com (L.F.); tankpow@aliyun.com (S.Q.)

[2] Guangzhou Institute of Energy Conversion, Chinese Academy of Sciences, Guangzhou 510640, China; yetong@mail.ustc.edu.cn (T.Y.); huangyp@ms.giec.ac.cn (Y.H.)

[3] School of Energy Science and Engineering, University of Science and Technology of China, Hefei 230026, China

[*] Correspondence: yangyuyao@gd.csg.cn

Abstract: As global energy demand rises and climate change poses an increasing threat, the development of sustainable, low-carbon energy solutions has become imperative. This study focuses on optimizing shared energy storage (SES) and distribution networks (DNs) using deep reinforcement learning (DRL) techniques to enhance operation and decision-making capability. An innovative dynamic carbon intensity calculation method is proposed, which more accurately calculates indirect carbon emissions of the power system through network topology in both spatial and temporal dimensions, thereby refining carbon responsibility allocation on the user side. Additionally, we integrate user-side SES and ladder-type carbon emission pricing into DN to create a low-carbon economic dispatch model. By framing the problem as a Markov decision process (MDP), we employ the DRL, specifically the deep deterministic policy gradient (DDPG) algorithm, enhanced with prioritized experience replay (PER) and orthogonal regularization (OR), to achieve both economic efficiency and environmental sustainability. The simulation results indicate that this method significantly reduces the operating costs and carbon emissions of DN. This study offers an innovative perspective on the synergistic optimization of SES with DN and provides a practical methodology for low-carbon economic dispatch in power systems.

Keywords: shared energy storage; ladder-type carbon price; low-carbon optimal scheduling; deep reinforcement learning; deep deterministic policy gradient algorithm

check for updates

Citation: Zhong, L.; Ye, T.; Yang, Y.; Pan, F.; Feng, L.; Qi, S.; Huang, Y. Deep Reinforcement Learning-Based Joint Low-Carbon Optimization for User-Side Shared Energy Storage–Distribution Networks. *Processes* **2024**, *12*, 1791. https://doi.org/10.3390/pr12091791

Academic Editors: Son Ich Ngo and Hoang Long Ngo

Received: 19 June 2024
Revised: 15 August 2024
Accepted: 21 August 2024
Published: 23 August 2024

1. Introduction

As global energy demand rises and climate change poses an increasing threat, the development of sustainable, low-carbon energy solutions has become imperative. According to the International Energy Agency (IEA) [1], global energy demand is expected to nearly double in the next two decades, while the latest assessment report of the United Nations Framework Convention on Climate Change (UNFCCC) highlights the need for significant reductions in greenhouse gas emissions in order to limit the rise in global temperature to more than 1.5 °C [2]. Direct carbon emissions refer to the greenhouse gases emitted by power generation companies during energy consumption and production processes, which can be directly controlled and managed. In contrast, indirect carbon emissions are those resulting from the energy (such as electricity, steam, heating, and cooling) purchased by companies. Although these emissions are not produced directly by the companies, they are intrinsically linked to their operations.

The power sector is the largest consumer of energy, making it crucial to control and reduce user-side indirect carbon emissions, particularly on the distribution network side. This not only helps to decrease the overall carbon footprint but also promotes the sustainable development of the energy supply chain. Low-carbon optimization refers to the systematic approach taking the objective of minimizing carbon emissions across various operational processes while maintaining or enhancing system performance. In the context of energy systems, this involves the optimization of energy production, distribution, and consumption to reduce both direct and indirect carbon emissions. Kang et al. [3] introduced a method for calculating indirect carbon emissions in power systems using a carbon emission flow model. This approach refines the calculation of carbon emission intensity to smaller spatial and temporal granularities, clarifying the concept of indirect carbon emission flow. Building on this, utilizing locational marginal prices and dynamic carbon intensity signals has demonstrated economic and environmental benefits within distribution networks (DN) [4]. Moreover, a low-carbon economic dispatch model enhances energy flexibility and decreases pressure in these networks [5]. Additionally, a tiered carbon emission model, addressing uncertainties in wind, photovoltaic generation, and loads, has been developed to minimize emissions and costs within a carbon trading framework [6]. Finally, a production simulation method that integrates diverse energy outputs with a carbon trading mechanism has also been proposed [7], paving the way for advances in energy system optimization.

By integrating advanced smart grid technology, renewable energy, and energy storage systems, DN can respond more flexibly to complex and variable energy demands [8,9]. This integration can not only promote the widespread use of renewable energy but also enhance the system's flexibility in responding to fluctuations in energy demand, thereby providing more reliable and efficient power services [10]. Applying various scheduling strategies and a dynamic carbon emission trading system, the efficient maximization of renewable energy for power generation and the reduction of carbon emissions have been achieved [11]. Mitigating indirect carbon emissions on the user side is crucial. However, research on user-side SES systems has not fully considered the application of dynamic carbon emission intensity. The importance of dynamic carbon emission intensity in storage operation strategies are not well addressed. Thus, this paper aims to explore the effective integration of dynamic carbon factors and carbon emission flow theory into user-side shared energy storage–distribution network systems, addressing the research gap in this critical area.

On the distribution network side, the application of energy storage systems, particularly within user-side shared energy storage–distribution grids, has proven effective in reducing user electricity costs and decreasing indirect carbon emissions. To enhance this, a proposal suggests utilizing energy storage efficiently and safely as a flexible grid asset by employing an energy management system (EMS) and optimization techniques to deliver various electric power services to users [12,13]. Furthermore, ref. [14] introduces a double-layer optimal distribution method for the cooperative optimization of distributed shared energy storage within networks, along with a corresponding operational model. Simulation analysis of numerical examples verifies the effectiveness and economic viability of this proposed configuration method.

Unlike traditional models that dedicate energy storage to individual users, "User-Side Shared Energy Storage-Distribution Grids" offer a modern approach by pooling storage assets to serve multiple users within a distribution network. This system is managed by a centralized EMS, which optimizes energy distribution based on real-time demand, pricing, and carbon emission signals [15,16]. This setup not only enhances system flexibility but also significantly improves the efficiency of energy storage utilization. By sharing storage costs and leveraging load complementarity among users, these systems effectively reduce the overall carbon footprint [17,18]. Shared energy storage (SES) plays a crucial role by assessing complementary storage capacities and proposing coordinated operation strategies to efficiently serve customers [19]. Additionally, using Nash bargaining theory, a dynamic zonation optimization strategy for centralized SES power stations has been developed, further enhancing storage utilization and benefiting various stakeholders [20].

Several existing studies have demonstrated that SES provides considerable advantages in terms of environmental economic benefits and practical engineering value [21,22] have shown that SES not only increases the flexibility and efficiency of the system, but also is a key technology to support the development of low-carbon power grids. This forward-looking mechanism not only promotes the integration of energy storage resources, but also increases the flexibility and efficiency of the system, making SES a key component to support low-carbon power grids. The development and implementation of low-carbon optimization strategies has become a necessary measure to reduce environmental impact, and SES shows great potential to address climate change by reducing carbon emissions [23]. A double-layer carbon-aware planning method for SES stations for multi-component integrated energy systems was proposed [24]. The conditional value-at-risk method was adopted as a risk measurement to effectively reduce the carbon emissions and system operating costs of SES stations. A power system interval optimization model based on SES and refined demand response is proposed to effectively deal with the uncertainty of source load through interval optimization and enhance the utilization rate of energy storage as well as the overall system economy [25].

However, traditional optimization algorithms struggle with high computational complexity, dynamic environments, and uncertainty in distribution network (DN) optimization problems [26]. With advancements in artificial intelligence, deep reinforcement learning (DRL) has emerged as a powerful tool for optimizing the operations of SES and DNs [27]. DRL's ability to learn and adapt through continuous interaction with complex, dynamic environments makes it particularly well-suited for DN optimization. This study specifically employs the deep deterministic policy gradient (DDPG) algorithm, a sophisticated DRL technique, to revolutionize optimization and decision-making processes within these systems. DDPG excels in handling complex optimization problems, adapting to dynamic environments, managing uncertainties, and being model-independent, thus providing SES and DN operators with highly flexible and efficient decision support tools [28,29]. As a model-free reinforcement learning algorithm, DDPG can independently learn and make decisions across successive actions and states [30]. Studies have shown that DRL agents outperform stochastic optimization algorithms in extensive action and observation spaces and can effectively manage uncertainty with high accuracy [31].

By integrating Lyapunov optimization strategies with DDPG, the challenges of real-time operation in wind-storage integrated systems (WSISs) are effectively addressed, particularly under conditions of uncertainty and fluctuating electricity prices [32]. Case studies demonstrate the effectiveness of DDPG in enhancing power system efficiency and reliability [33]. Additionally, a reward function based on the long-term behavior of energy systems enables optimization of online scheduling by learning the operational modes of renewable energy generation and grid dynamics [34]. This demonstrates that DRL not only enhances SES efficiency but also supports the low-carbon transformation of DNs, offering significant economic and environmental benefits.

Traditional power grids face challenges in achieving environmentally friendly, low-carbon operations while maintaining a stable power supply. [35].Although previous studies have utilized various advanced deep reinforcement learning algorithms to optimize power system scheduling, few have simultaneously considered dynamic carbon emission intensity, photovoltaic generation uncertainty, and shared energy storage for low-carbon economic dispatch. To achieve the dual goals of economic efficiency and environmental sustainability in energy systems, this study aims to apply the DDPG algorithm with prioritized experience replay and orthogonal regularization to achieve low-carbon optimization of shared energy storage and DN under the conditions of uncertain carbon intensity and photovoltaic output, ensuring real-time adaptability to fluctuating carbon intensities and renewable energy outputs.

The main contributions of this study are as follows:

1. To address the uncertainties in carbon emission intensity of the main grid and the photovoltaic generation output within the DN, a low-carbon optimization model is

proposed. This model incorporates tiered carbon trading and considers the dynamic carbon emission intensity, aiming to optimize the operation of SES operators in conjunction with the DN.

2. Based on the aforementioned model, the DDPG algorithm with prioritized experience replay and orthogonal regularization is employed, which improves the convergence and stability of the algorithm. This approach ensures real-time adaptability to fluctuating carbon intensities and renewable energy outputs.

2. Problem Description and Framework

The framework of this study is illustrated in Figure 1. Within the DN, various distributed energy resources, energy storage systems, and photovoltaic (PV) installations are distributed. The battery energy storage systems (BESS) in user microgrids are managed by a SES operator, ensuring efficient utilization and management of energy resources. All users within the DN participate in the carbon emission rights trading market, leveraging market mechanisms to optimize and control carbon emissions, thus promoting low-carbon development goals.

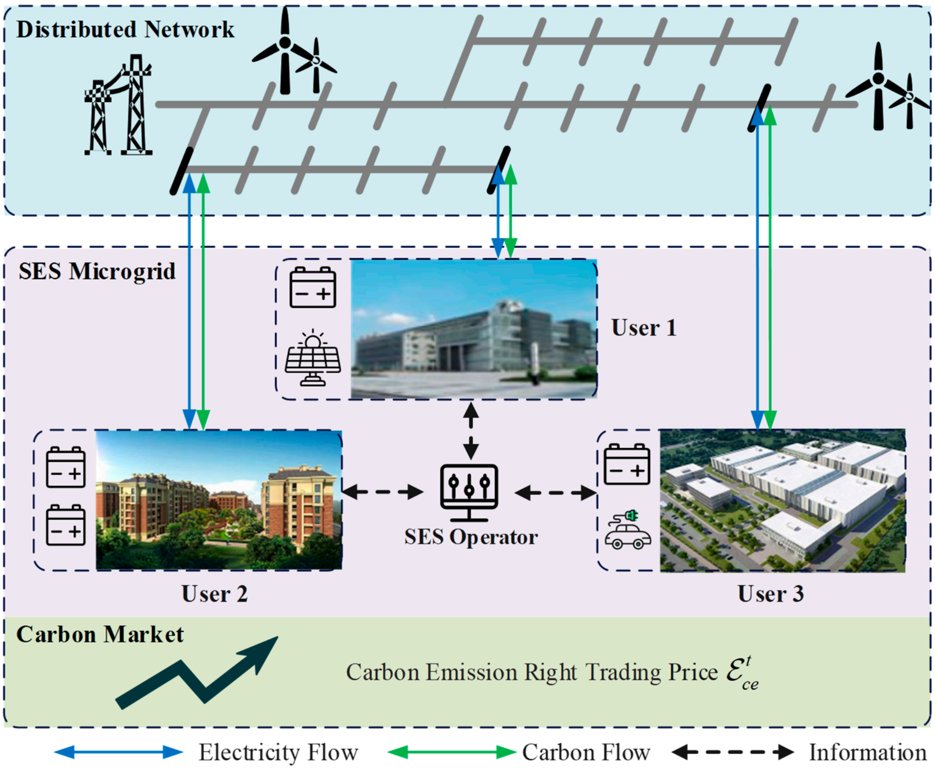

Figure 1. Shared energy storage–distribution system framework.

Notably, the DN is connected to the main grid, which provides time-segmented carbon emission intensity data. Combined with the internal energy structure of the DN, the carbon emission intensity within the entire DN dynamically changes. This dynamic characteristic allows the DN to flexibly adjust its operational strategies, optimize carbon emissions, and maximize the use of low-carbon energy at different times, thereby achieving more efficient carbon emission management and energy utilization.

2.1. Distribution Network Modeling

The objective function of the economic low-carbon dispatch problem is demonstrated in (1), aiming to minimize the total cost. This objective function consists of four terms, including the cost of power purchased from the higher grid by distribution grid users $C_{E,i}^{t}$, the cost of gas turbine operation by microgrid users in the system $C_{E,i}^{t}$, the profit of the SES service provider $C_{k,t}^{BESS}$, and the cost of carbon trading for the system segmented carbon price C_{ce}^{t}, as detailed in (14).

$$\begin{aligned}\min F &= \sum_{t=1}^{T}\left(\sum_{i\in\Omega^{ADN}} C_{i,t}^{E} + \sum_{i\in\Omega^{G}} C_{i,t}^{G} + \sum_{k\in\Omega^{BESS}} C_{k,t}^{BESS} + C_{t}^{ce}\right) \\ &= \sum_{t=1}^{T}\left(\begin{array}{l} \sum\limits_{i\in\Omega^{ADN}} \lambda_t^{ELE} P_{i,t}^{l}\Delta t + \sum\limits_{g\in\Omega^{G}} \left(a_i (P_{g,t}^{G}\Delta t)^2 + b_i P_{g,t}^{G}\Delta t + c_i\right) \\ + \sum\limits_{k\in\Omega^{BESS}} \left(\begin{array}{c} (\lambda_t^{ELE} P_{k,t}^{ch} - \lambda_t^{ELE}\left|P_{k,t}^{dis}\right|)\Delta t \\ -\lambda^{BESS}(P_{k,t}^{ch} + \left|P_{k,t}^{dis}\right|)\Delta t \end{array}\right) + C_{ce}^{t} \end{array}\right)\end{aligned} \tag{1}$$

where Ω^{ADN} is the set of all power buses in the system; Ω^{G} is the set of consumer gas turbines in the system; Ω^{BESS} is the set of SES; λ_t^{ELE} is the time-of-use tariffs; $P_{i,t}^{l}$ is the nodal consumer loads; $P_{g,t}^{G}$ is the generator outputs; a_i, b_i, and c_i are the cost coefficients; $P_{k,t}^{ch}$ is the charging power of SES; $P_{k,t}^{dis}$ is the discharging power of SES; and λ^{BESS} is the cost of operation of SES, taking into account the lifetime of the batteries.

The constraints are established as follows.

$$\begin{aligned}&\sum_{i\in\Omega^{ADN}} P_{i,t}^{l} - \sum_{i,j\in\Omega^{ADN}} (P_{ij,t} - r_{ij} l_{ij}) + \sum_{k\in\Omega^{BESS}} P_{k,t}^{ch} \\ &= P_t^{UPG} + \sum_{p\in\Omega^{PV}} P_{p,t}^{PV} + \sum_{g\in\Omega^{G}} P_{g,t}^{G} + \sum_{k\in\Omega^{BESS}} \left|P_{k,t}^{dis}\right|\end{aligned} \tag{2}$$

$$\sum_{i\in\Omega^{ADN}} Q_{i,t}^{l} - \sum_{i,j\in\Omega^{ADN}} (Q_{ij,t} - x_{ij} l_{ij}) = Q_t^{UPG} + \sum_{p\in\Omega^{PV}} Q_{p,t}^{PV} + \sum_{g\in\Omega^{G}} Q_{g,t}^{G} \tag{3}$$

$$\sqrt{(P_{ij,t})^2 + (Q_{ij,t})^2} \le S_{ij}^{\max}, \forall i,j \in \Omega^{ADN} \tag{4}$$

$$v_{j,t} = v_{i,t} + \left(r_{ij}^2 + x_{ij}^2\right) l_{ij,t} - 2\left(r_{ij} P_{ij,t} + x_{ij} Q_{ij,t}\right), \forall i,j \in \Omega^{ADN} \tag{5}$$

$$\left(V_i^{\min}\right)^2 \le v_{i,t} \le (V_i^{\max})^2, \forall i \in \Omega^{ADN} \tag{6}$$

$$\left(I_{ij}^{\min}\right)^2 \le l_{ij,t} \le \left(I_{ij}^{\max}\right)^2, \forall i,j \in \Omega^{ADN} \tag{7}$$

$$P_k^{ch,\min} \le P_{k,t}^{ch} \le P_k^{ch,\max}, \forall k \in \Omega^{BESS} \tag{8}$$

$$P_k^{dis,\min} \le P_{k,t}^{dis} \le P_k^{dis,\max}, \forall k \in \Omega^{BESS} \tag{9}$$

$$P_g^{G,\min} \le P_{g,t}^{G} \le P_g^{G,\max}, \forall g \in \Omega^{G} \tag{10}$$

$$E_{k,t+1}^{BESS} = E_{k,t}^{BESS} + \eta^{ch} P_{k,t}^{ch}\Delta t - \frac{P_{k,t}^{dis}}{\eta^{dis}}\Delta t, \forall k \in \Omega^{BESS} \tag{11}$$

$$E_k^{BESS,\min} \le E_{k,t}^{BESS} \le E_k^{BESS,\max}, \forall k \in \Omega^{BESS} \tag{12}$$

where $P_{ij,t}$ is the power flowing through the branch; r_{ij} is the branch resistance; l_{ij} is the square of the current flowing through the branch; P_t^{UPG} is the inflow power from the higher grid; $P_{p,t}^{PV}$ is the active power output from the PV inverter; $Q_{i,t}^{l}$ is the reactive power load at the bus; $Q_{ij,t}$ is the reactive power flowing through the branch; Q_t^{UPG} is the reactive power inflow from the higher grid; $Q_{p,t}^{PV}$ is the reactive power output from the PV inverter; $Q_{g,t}^{G}$ is the reactive power output from the user's micro-gas turbine; $S_{ij}^{\max}$ is the maximum capacity of the line; $v_{i,t}$ is the square of the bus voltage; $V_i^{\min}$ and $V_i^{\max}$ are the minimum and maximum values of the bus voltage; $I_{ij}^{\min}$ and $I_{ij}^{\max}$ are the minimum and maximum values of the branch current; $P_k^{ch,\min}$ and $P_k^{ch,\max}$ are the minimum and maximum values of the charging power; $P_k^{dis,\min}$ and $P_k^{dis,\max}$ are the minimum and maximum values of the discharging power; $P_g^{G,\min}$ and $P_g^{G,\max}$ are the minimum and maximum values of the microgrid users' gas turbine output.

Equations (2) and (3) denote the system's active and reactive power balance equations, respectively, ensuring that the total generated active and reactive power equals the total consumed active and reactive power within the distribution network. This balance is crucial for maintaining system stability. Equation (4) denotes the line capacity constraint, ensuring that the power flow through each line does not exceed its maximum capacity. Equation (5) denotes the bus voltage balance, ensuring that the voltage at each bus remains within acceptable limits. Equation (6) represents the bus voltage magnitude constraint, ensuring that the voltage magnitude at each bus stays within specified bounds. Equation (7) denotes the branch current magnitude constraint, ensuring that the current in each branch does not exceed its maximum limit. Equation (8) denotes the energy storage charging constraint, which ensures that the charging power of energy storage systems remains within their capacity limits. Equation (9) denotes the energy storage discharge constraint, ensuring that the discharge power of energy storage systems does not exceed their capacity. Equation (10) denotes the user microgrid gas turbine output constraint, which limits the power output of gas turbines in microgrids. Equation (11) denotes the relationship between the capacity of the energy storage during charging and discharging processes, ensuring proper coordination between the two. Finally, Equation (12) denotes the capacity constraint of the BESS, ensuring that the energy storage operates within its designated capacity.

2.2. System Carbon Intensity Calculation

Dynamic carbon emission intensity varies based on the proportion of different energy sources utilized within the grid, fluctuations in load, and the generation capacity of renewable energy sources. Unlike static carbon factors, dynamic carbon emission intensity provides real-time reflections of the environmental impact of electricity supply, thereby offering users more accurate data on carbon emissions. The carbon intensity of the system is given by Equation (13):

$$C_t = \frac{P_t^{UPG} C_t^{UPG} + \sum\limits_{g \in \Omega^G} P_{g,t}^G C^G}{P_t^{UPG} + \sum\limits_{g \in \Omega^G} P_{g,t}^G + \sum\limits_{p \in \Omega^{PV}} P_{p,t}^{PV} + \sum\limits_{k \in \Omega^{BESS}} \left| P_{k,t}^{dis} \right|} \tag{13}$$

where C_t^{UPG} is the real-time dynamic carbon intensity of the superior grid and C^G is the carbon intensity of the gas turbine of the user microgrid.

2.3. Ladder-Type Carbon Emission Right Trading

Simulating a real carbon emission market trading scenario is complex; hence, the ladder-type carbon trading approach, which is widely accepted [36,37], is used to mimic real-world carbon trading scenarios. This method follows the market principle that the scarcer the commodity, the higher the price, while maintaining a constant total quantity. The total carbon emissions $\Delta Q = \sum\limits_{t=1}^{T} (C_t(\sum\limits_{i \in \Omega^{ADN}} P_{i,t}^l + \sum\limits_{k \in \Omega^{BESS}} P_{k,t}^{ch}))$ are calculated at the end of all scheduling periods of the day, and then the growth interval d and the growth rate of the carbon trading cost λ are determined based on the historical data. The higher the carbon emissions, the higher the corresponding carbon trading price. Equation (14) denotes the ladder-type carbon emission trading cost.

$$C_{ce}^t = \begin{cases} \varepsilon_{ce}^t \Delta Q, 0 < \Delta Q < d \\ \varepsilon_{ce}^t d + (1+\lambda)\varepsilon_{ce}^t (\Delta Q - d), d < \Delta Q < 2d \\ \varepsilon_{ce}^t d + (1+\lambda)\varepsilon_{ce}^t d + (1+2\lambda)\varepsilon_{ce}^t (\Delta Q - 2d), 2d < \Delta Q < 3d \\ \varepsilon_{ce}^t d + (1+\lambda)\varepsilon_{ce}^t d + (1+2\lambda)\varepsilon_{ce}^t d + (1+3\lambda)\varepsilon_{ce}^t (\Delta Q - 3d), \Delta Q > 3d \end{cases} \tag{14}$$

3. Markov Decision Process Framework of Dynamic Dispatch

In this study, deep reinforcement learning is utilized to solve the low-carbon economic dispatch problem of a distribution grid–SES system. First, the mathematical formulation of this low carbon economy scheduling problem is transformed into the MDP framework for reinforcement learning. The design of state space, action space, and reward function is included.

A Markov decision process (MDP) [38] is a mathematical model used to describe decision-making in environments that involve randomness and decision-making. In reinforcement learning and decision theory, MDP provides a systematic framework that allows complex decision problems to be solved through mathematical and algorithmic methods. An MDP typically consists of the following four components: State Set (S): A set representing all possible states the system can be in. At any given time, the system is in a specific state. Action Set (A): A set representing all possible actions. The agent (decision-maker) can choose an action $a \in A$ in each state s. State Transition Probability (P): This describes the probability of transitioning from one state to another. Specifically, $P(s' \mid s, a)$ represents the probability of transitioning to state s' after taking action a in state s. Reward Function (R): This describes the immediate reward received after taking a specific action in a specific state. It is usually represented as $R(s, a)$, which is the reward received after taking action a in state s. The MDP framework in our study is designed to optimize the low-carbon economic dispatch of the SES–distribution grid system. The state space includes variables such as electrical load, energy prices, carbon emissions, and the state of charge of the SES. The action space comprises the control actions for the SES and microgrid generators. The reward function is designed to minimize the total operating cost and carbon emissions, incorporating penalties for violating constraints.

The agent obtains state observations s_t through interaction with the actual environment, and these observations provide key data for effective training of the agent based on the perceived state information. The study utilizes a 24 h time frame, during which the system observations collected include electrical load data at a specific time period (period t), energy prices, carbon emissions trading volume, tiered carbon prices, and the state of charge of the BESS system. For this problem, the state can be expressed as

$$s_t = \left\{ P_{i,t}^{l}, P_{ij,t}, \varepsilon_{ce}^{t}, \lambda_t^{ELE}, SOC_{k,t-1}^{BESS}, \Delta Q \right\}, \forall k \in \Omega^K \tag{15}$$

The scheduling objective is to determine the optimal SES as well as PV and wind generation outage scenarios, and the set of actions is denoted as A_i, $a_i \in A_i$. Actions in low-carbon scheduling can be expressed as follows:

$$a_t = \left\{ P_{g,t}^{CHP}, P_{k,t}^{BESS} \right\}, \forall g \in \Omega^G, \forall k \in \Omega^K \tag{16}$$

where $P_{g,t}^{CHP} = \left\{ P_{g1,t}^{CHP}, P_{g2,t}^{CHP}, P_{g3,t}^{CHP} \right\}$ denotes the action of each microgrid user's gas turbine, and $P_{k,t}^{BESS} = \left\{ P_{k1,t}^{BESS}, P_{k2,t}^{BESS}, P_{k3,t}^{BESS}, P_{k4,t}^{BESS}, P_{k5,t}^{BESS} \right\}$ denotes the action of each storage battery for SES. A is the continuous action space that satisfies the energy and capacity constraints that satisfy the action constraints.

The state transfer is $p\left(s_{t+1} \middle| s_t, \mathcal{A}^t\right)$, and in any time step, state s_t and action a_t for a given time step t are transitioned to the next state s_{t+1} via the distribution grid–SES.

The reward received by the intelligent agent during the time period t is given by the environment in order to guide the intelligent agent to update its strategy to minimize the carbon trading price and operation cost to achieve the goal of low-carbon economy operation by designing a suitable reward function. After the agent makes an action and interacts with the environment, the environment gives it a reward value.

$$
\begin{aligned}
r_t = & \sum_{i\in\Omega^{ADN}} \lambda_t^{ELE} P_{i,t}^{l} + \sum_{g\in\Omega^{G}} \left(a_i (P_{g,t}^{G})^2 + b_i P_{g,t}^{G} + c_i \right) + C_{ce}^{t} \\
& + \sum_{k\in\Omega^{BESS}} \left(\begin{array}{l} (\lambda_t^{ELE} P_{k,t}^{ch} - \lambda_t^{ELE} \left| P_{k,t}^{dis} \right|) \\ -\lambda^{BESS} (P_{k,t}^{ch} + \left| P_{k,t}^{dis} \right|) \end{array} \right) + (\lambda_1 E_{k,t}^{BESS} + \lambda_2 P_{g,t}^{G})
\end{aligned} \tag{17}
$$

The reward function contains five terms: the first four terms correspond to the objective function of the modeling approach in Section 2 as Equation (1), C_{ce}^{t} is added to the reward in the last step of each episode as the settlement of daily carbon emissions trading, and the last term is the penalty term for constraint Equations (10) and (12).

4. DDPG Algorithm for Online Dispatch Problem

This study uses the DDPG algorithm enhanced with prioritized experience replay (PER) and orthogonal regularization (OR) to solve the above MDP problem, and works to learn the optimal economic low-carbon green scheduling strategy. DDPG combines the advantages of deep learning (DL) and reinforcement learning (RL), and is particularly suitable for dealing with problems in continuous action space. The algorithm mainly consists of two parts: the strategy network $\mu(s|\theta^{\mu})$ and its corresponding goal network $\mu(s|\theta^{\mu\prime})$; and the value network $Q(s_t, \mu(s_t)|\theta^{Q})$ and its corresponding goal network $Q(s_t, \mu(s_t)|\theta^{Q\prime})$.

The policy network generates the action a_t based on the current state of the environment s_t, while the value network evaluates the expected payoff of this action for taking a particular action under the current policy. Thus, the strategy network is trained to maximize the action evaluation given by the value network:

$$
\max_{\theta^{\mu}} J(\theta^{\mu}) \tag{18}
$$

where $J(\theta^{\mu}) = \mathbb{E}_{s_t\sim\rho^{\beta}}[Q(s,a\left|\theta^{Q}\right)\Big|_{s=s_t,a=\mu(s_t|\theta^{\mu})}]$

The gradient rise is utilized to increase $J(\theta^{\mu})$, as shown in Equation (19):

$$
\begin{aligned}
\nabla_{\theta^{\mu}} J &\approx \mathbb{E}_{s_t\sim\rho^{\beta}} \left[\nabla_{\theta^{\mu}} Q(s,a\left|\theta^{Q}\right)\big|_{s=s_t,a=\mu(s_t|\theta^{\mu})} \right] \\
&= \mathbb{E}_{s_t\sim\rho^{\beta}} \left[\nabla_a Q(s,a\left|\theta^{Q}\right)\Big|_{s=s_t,a=\mu(s_t)} \nabla_{\theta_{\mu}} \mu(s\left|\theta^{\mu}\right)\Big|_{s=s_t} \right]
\end{aligned} \tag{19}
$$

Thus, the algorithm for updating the policy network θ^{μ} is obtained by randomly drawing one state at a time from the replay buffer, denoted as s_j, computing $\hat{a}_j = \mu(s_j|\theta^{\mu})$, and updating θ^{μ} once with gradient ascent, as shown in Equation (20).

$$
\theta^{\mu} \leftarrow \theta^{\mu} + \beta \cdot \nabla_a Q(s,a\left|\theta^{Q}\right)\Big|_{s=s_t,a=\mu(s_t)} \nabla_{\theta_{\mu}} \mu(s\left|\theta^{\mu}\right)\Big|_{s=s_t} \tag{20}
$$

The value network $Q(s_t, \mu(s_t)|\theta^{Q})$ learns to update the parameters θ^{Q} by temporal difference (TD) learning errors, and each time, a quaternion (s^j, a^j, r^j, s^{j+1}) from the replay buffer computes the predicted values $Q(s_t, a_t|\theta^{Q})$ and $Q(s_{t+1}, \mu(s_{t+1})|\theta^{Q})$ of the target value network, and then computes the TD target as shown in Equation (21).

$$
y_t = r(s_t, a_t) + \gamma Q(s_{t+1}, \mu(s_{t+1})\left|\theta^{Q}\right) \tag{21}
$$

The loss function is defined as shown in Equation (22):

$$
L(\theta^{Q}) = \frac{1}{2}\left(Q(s_t, a_t\left|\theta^{Q}\right) - y_t \right)^2 \tag{22}
$$

The computed gradient is used to update the parameters θ^Q of the value network as shown in Equation (23):

$$\nabla_{\theta^Q} L(\theta^Q) = (Q(s_t, a_t \middle| \theta^Q) - y_t)\nabla_{\theta^Q} Q(s_t, a_t \middle| \theta^Q) \tag{23}$$

The parameter updates for the strategy network are calculated as shown in Equation (24):

$$\theta^Q \leftarrow \theta^Q - \alpha \cdot (Q(s_t, a_t \middle| \theta^Q) - y_t)\nabla_{\theta^Q} Q(s_t, a_t \middle| \theta^Q) \tag{24}$$

In order to improve the stability and convergence of training, the parameters of the two networks are updated using soft update, for the target strategy network and target value network, the parameters are updated as shown in Equations (25) and (26).

$$\theta^{\mu'} \leftarrow \tau \cdot \theta^{\mu} + (1 - \tau) \cdot \theta^{\mu'} \tag{25}$$

$$\theta^{Q'} \leftarrow \tau \cdot \theta^{Q} + (1 - \tau) \cdot \theta^{Q'} \tag{26}$$

where $\tau << 1$ is the soft update rate, which reduces the impact of abrupt changes in network parameters on the training process by slowly adjusting the parameters of the target network to gradually converge to those of the actual network.

Prioritized experience replay (PER) [39] enhances the efficiency of reinforcement learning by selectively sampling experiences that are more informative. Unlike uniform sampling in traditional experience replay, PER assigns higher sampling probabilities to experiences with larger temporal difference (TD) errors. This focus on high-priority experiences accelerates learning and stabilizes the training process. In our study, PER is integrated into the DDPG algorithm, which helps the model learn more effectively from the most significant experiences. The priority of each experience is given by Equation (27):

$$p_i = (\delta_i + \epsilon)^{\alpha} \tag{27}$$

Orthogonal regularization (OR) [40] is employed to improve the generalization and stability of neural networks by maintaining the orthogonality of weight matrices. This method helps prevent overfitting and enhances the robustness of the learned policies. By incorporating an orthogonal regularization term into the loss function, we encourage the weight matrices to remain orthogonal, thus improving the performance of deep learning models. In our implementation, OR is added to the DDPG algorithm, ensuring that the neural network weights are regularized throughout training. The orthogonal regularization term is defined as Equation (28):

$$L_{or} = \frac{\lambda}{2} \sum_i \parallel W_i^T W_i - I \parallel_F^2 \tag{28}$$

where W_i is the weight matrix of the i-th layer, I is the identity matrix, λ is the regularization coefficient, and $\| \cdot \|_F$ denotes the Frobenius norm. This addition helps stabilize the learning process and enhances the algorithm's ability to handle uncertainties in low-carbon energy dispatch scenarios.

When the DDPG agent completes the offline training, only the trained policy network is used to make decisions, and the Agent can make online decisions through its own observation, which is based on the state in the observation area, and online decisions at the millisecond level are made through the already trained policy network. The overall block diagram of the algorithm based on DDPG for the joint low-carbon optimization of the user-side SES–distribution network is shown in Figure 2.

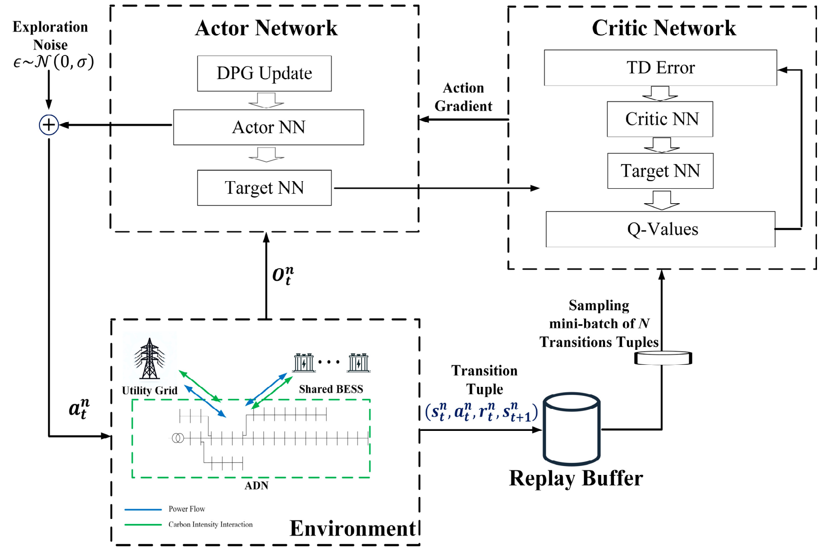

Figure 2. Block diagram of DDPG algorithm.

5. Case Analysis

5.1. Simulation Experimental Settings

This study validates the effectiveness of the proposed joint user-side SES-distribution grid low-carbon optimized online scheduling on a modified IEEE 33-bus distribution grid. A total of five grid-connected SES charging stations, three user microgrid gas turbines, and five PV power generation plants configured with inverters are located on the distribution grid, as shown in Figure 3. The load data of each bus are shown in Figure 4.

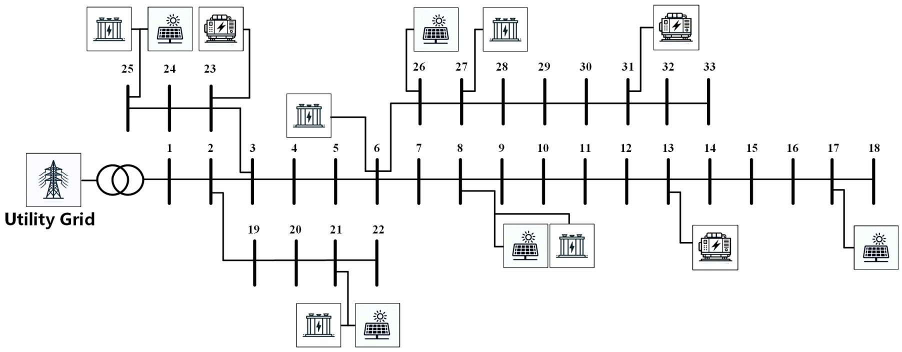

Figure 3. The modified IEEE-33 bus test system.

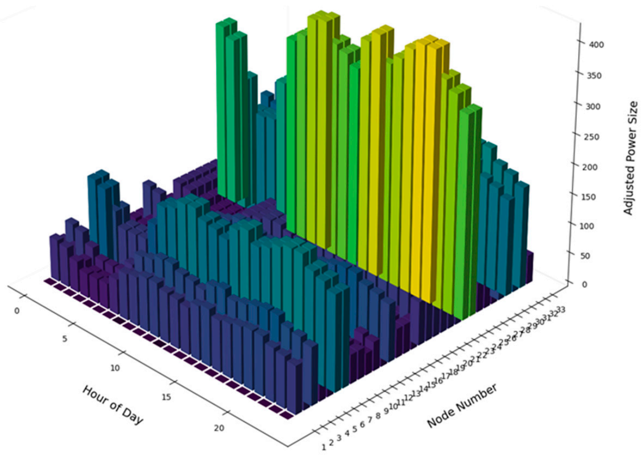

Figure 4. Twenty-four-hour load variation at each bus.

In this study, the RL algorithm is used to train and learn the low-carbon economy strategy, and the training is based on Python 3.6, Pytorch 2.1.2, Pandapower 2.13.1, and Gymnasium 0.29.1.

The 24 h carbon intensity is derived from the data released by a real-time carbon emission platform in a province in China, and the electricity price is set as a one-day time-of-day tariff in a region. In the DN scenario, the adjustment range for the energy storage system and micro-gas turbine are set to $[-300, 300]$ kW and $[0, 300]$ kW, respectively (Tables 1 and 2).

Table 1. Settings of the hyperparameter.

Hyperparameter	Value	Hyperparameter	Value
λ^G	\$500/MWh	η^{ch}	0.95
C^G	0.55 t/MWh	η^{dis}	0.8

Table 2. Settings of the proposed algorithm.

Hyperparameter	Value	Hyperparameter	Value
Total training episode	2×10^4	Steps/episode	24
Learning rate (actor)	4×10^{-4}	τ	0.01
Learning rate (critic)	4×10^{-4}	Gamma	0.96
Hidden layer dimension	256	Buffer size	10^5
Orthogonal initialization	✓	Prioritized experience replay	✓
Network	Structure		
Actor	Linear layer → layer normalization → ReLU → linear layer → ReLU → linear layer → tanh		
Critic	Linear layer → layer normalization → ReLU →linear layer → ReLU → linear layer		

5.2. Analysis of Offline Training Effect

During offline training, Figure 5 illustrates the reward curve of the system's integrated reward using both the traditional DDPG and the improved DDPG (proposed) algorithm. The results show that in the first 10,000 episodes, the agents are in the exploratory stage, with significant fluctuations in rewards as they learn to navigate the action space. After approximately 10,000 episodes, the rewards of the agents begin to stabilize, indicating that they have learned effective policies for the given tasks.

The blue curve represents the proposed DDPG algorithm, enhanced with prioritized experience replay (PER) and orthogonal regularization (OR), while the orange curve represents the traditional DDPG algorithm. The improved DDPG (proposed) algorithm demonstrates the same rapid convergence and higher final rewards compared to the traditional DDPG. The remaining fluctuations after stabilization are primarily due to the inherent randomness of the exploration noise and the uncertainty imposed by the system during the training process. These results underscore the success of the proposed improvements in boosting the learning efficiency and robustness of the DDPG algorithm, which incorporates PER and OR.

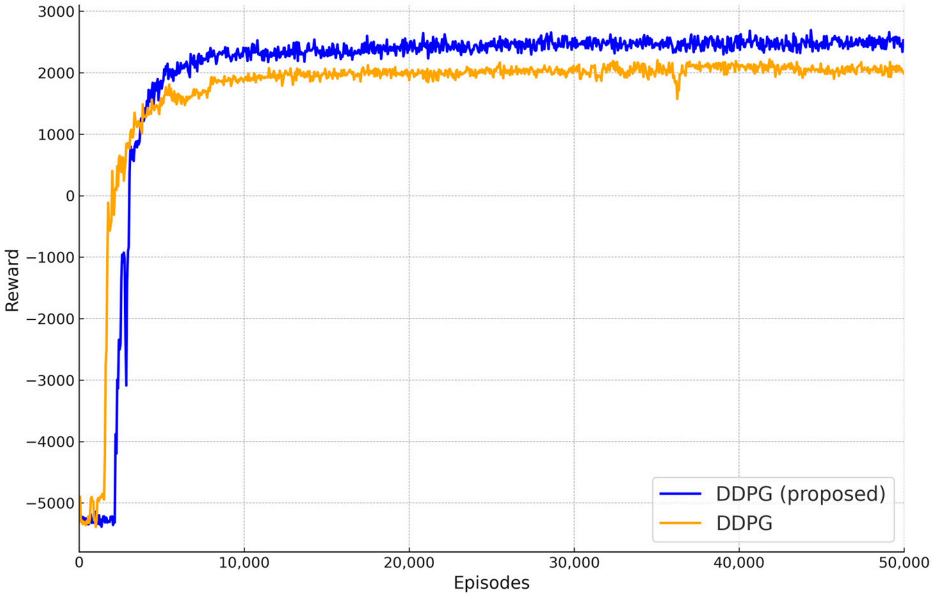

Figure 5. Training curve of DDPG algorithm.

5.3. DDPG Online Decision Effect Analysis

The agent's strategy network, trained for online decision making, is applied, and Figure 6 illustrates the carbon intensity of the main grid and the SES–distribution grid system at various times of the day, along with the actual emission reductions per hour. The figure reveals significant carbon reduction effects from 6:00 to 19:00, primarily due to the introduction of photovoltaic systems and the SES releasing stored energy during periods of lower carbon intensity. The orange line indicates the carbon intensity of the main grid, which remains relatively stable throughout the day, fluctuating between 0.5 and 0.6 t/MWh. The blue line represents the carbon intensity of the SES–distribution system, showing more variability compared to the main grid, with significant dips during certain periods (e.g., around 5:00–6:00 and 15:00–16:00). These dips coincide with periods when the SES likely discharges stored energy, reducing reliance on the main grid and thus lowering carbon intensity.

Notably, our carbon intensity reduction is calculated based on the difference between the main grid's carbon intensity and the DN's carbon intensity, multiplied by the total load at that time. Therefore, it can be seen that when the distribution grid's carbon intensity is lower (e.g., 8:00–15:00), the overall emission reduction of the system is considerable. From another perspective, the high proportion of clean energy in the distribution grid during these periods results in very low overall carbon intensity, significantly reducing indirect carbon emissions.

Moreover, the system's agent control strategy not only leverages clean energy during periods of low carbon emissions but also effectively avoids dependence on the main grid during high carbon emission periods (e.g., after 19:00), thereby achieving further emission reduction goals. This strategy not only enhances energy utilization efficiency but also significantly reduces overall carbon emissions. Further analysis of the data in Figure 6 reveals that during nighttime periods (e.g., 21:00–24:00), although the carbon intensity of the main grid is lower, the demand is relatively low, allowing the SES to continue providing power, thereby reducing reliance on the main grid and preventing potential increases in carbon emissions. This indicates that the optimized scheduling of the SES plays a crucial role in achieving low-carbon operation around the clock.

Figure 7 mainly shows the operation of the SES system under a time-sharing tariff. It can be seen that in the lower-tariff phase (0:00–7:00, 12:00–13:00), the agent controls the SES system to charge, and in the higher-tariff phase (9:00–11:00, 14:00–23:00), the agent controls the SES system to discharge, and due to the existence of the soft constraint of power, the state of charge (SOC) is maintained in the interval of 0.2–0.8; the SOC is maintained below 0.8 during charging and above 0.2 during discharging.

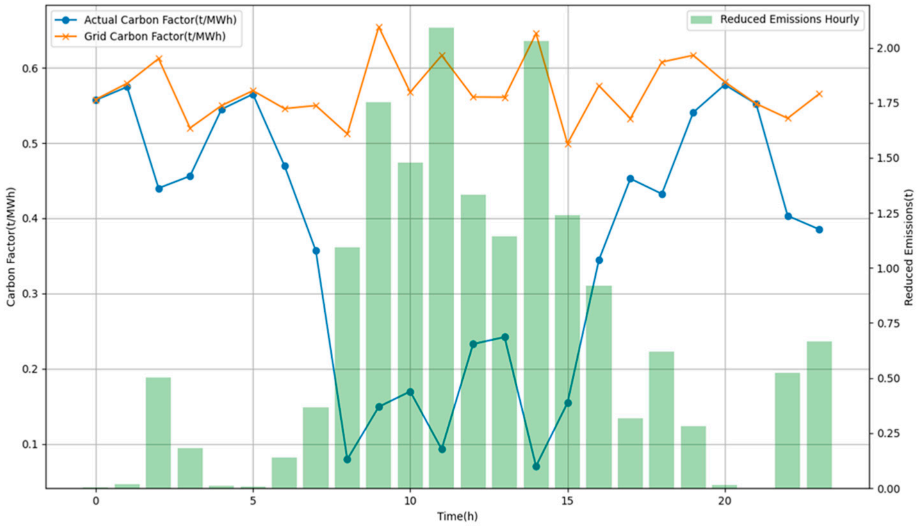

Figure 6. Carbon intensity and actual emission reductions of the main grid DN by time period.

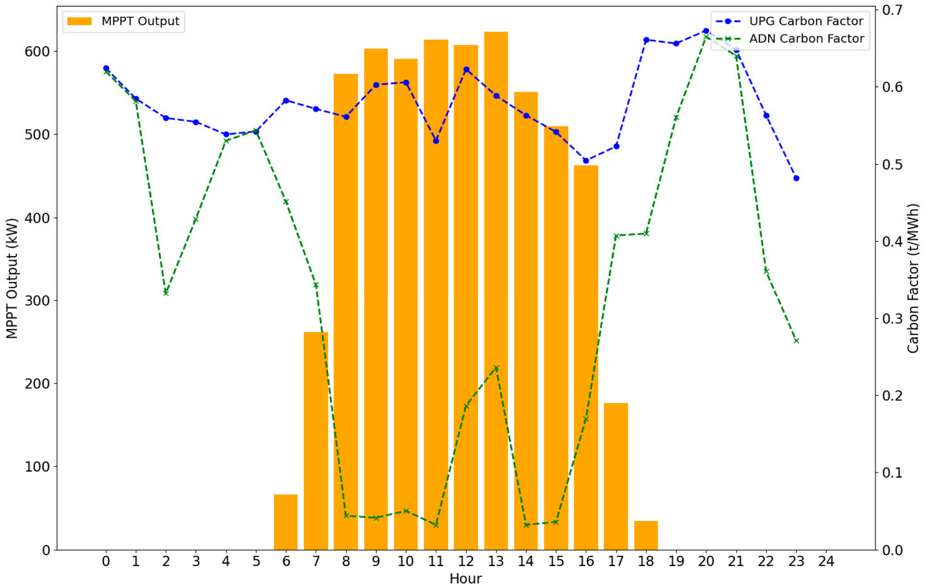

Figure 7. Correspondence between the carbon intensity of the main grid DN and the regional PV generation in each time period.

In fact, it is worth noting that Figure 7 shows the presence of energy storage #4 with an SOC above 0.8 at 7:00, as well as an SOC below 0.2 at 11:00 and 23:00. It occurs due to a combination of carbon intensity and electricity price. At these moments, the system exceeds the soft constraint because the profit from exceeding the constraint is greater than the amount of penalty set by the soft constraint. Therefore, the system chooses at some moments to exceed the operational lifetime constraint of the energy storage in order to obtain a correspondingly high profit.

The main focus of Figure 8 is to illustrate the substantial reduction in carbon intensity achieved by the system through the implementation of SES. By comparing it with Figure 9, a more comprehensive understanding can be gained regarding the pivotal role played by the SES system in mitigating carbon emissions during specific time periods where PV energy does not contribute (0:00–5:00 and 19:00–24:00). Of these time periods, of particular interest are 2:00–3:00 and 19:00–23:00, where the system demonstrates the clear carbon reduction effect of energy storage. The realization of this carbon reduction effect can be traced back to the intelligent scheduling strategy of the SES system. When PV energy is unavailable, the energy storage system effectively stores electricity by charging during periods of low grid demand (e.g., early morning hours). During periods of peak grid demand, especially in the evening, the SES system intelligently releases the stored power, thereby reducing the overall carbon intensity of the system (Table 3).

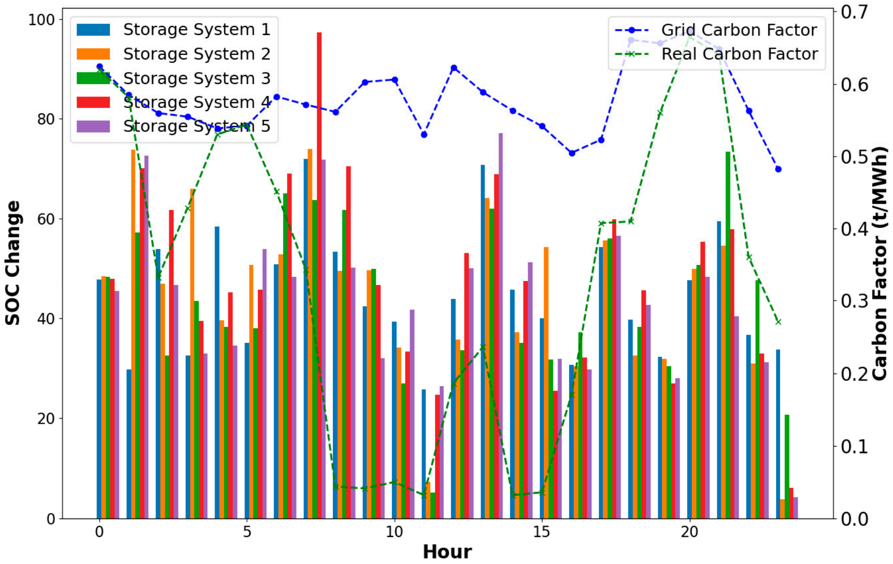

Figure 8. Correspondence between the carbon intensity of the main DN and the change in the state of charge of the SES for each time period.

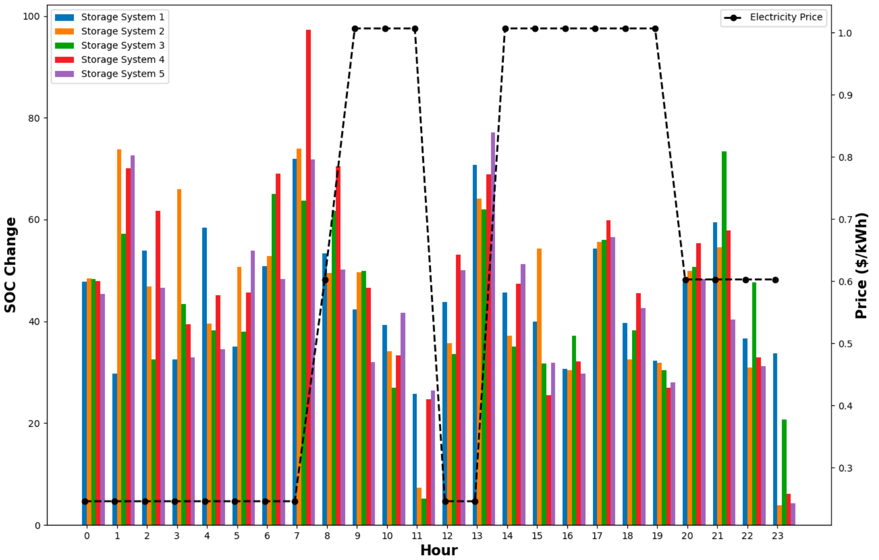

Figure 9. SES operation under time-of-use tariffs.

Table 3. Algorithm results.

Algorithm	Result		
	Target Convergence	Result Training Time	Decision Time
DDPG (proposed)	2561.656	22 h 26 min	0.678 s
DDPG	2009.749	8 h 37 min	0.683 s

6. Conclusions

This study is dedicated to exploring optimization strategies to reduce DN costs and carbon emissions. We address the system uncertainty caused by the high percentage of renewable energy access and the variability of the dynamic carbon factor in the main grid. By incorporating real-time dynamic carbon factors, our approach enables more accurate assessments of carbon emissions associated with electricity consumption, facilitating timely adjustments in energy dispatch and consumption patterns. To this end, we propose a joint SES operator–distribution grid low-carbon optimization model that considers ladder-type carbon trading and utilizes the DDPG algorithm enhanced with PER and OR to achieve the dual goals of economic efficiency and environmental sustainability.

We first propose an optimization framework that considers SES and all-day carbon trading costs to minimize operating costs and carbon emissions. Second, by formulating the scheduling problem as a MDP, the observed states, scheduling actions and reward functions of the system are explicitly defined. Finally, the DDPG algorithm is used for low-carbon economy scheduling. The simulation results of this study validate the proposed method's effectiveness in reducing both the operating cost of the power system and carbon emissions.

This study provides an innovative perspective to synergistically optimize SES with the distribution grid, and also offers a practical methodology for achieving low-carbon economic dispatch of the power system. Through a series of simulation experiments, the study confirms the significant effectiveness of the proposed methodology in reducing power system operating costs and reducing carbon emissions, providing an innovative perspective and a practical methodology for the low-carbon optimization of SES with distribution grids.

Future research directions include investigating the scalability of the proposed low-carbon optimization model in larger, more complex DNs and its deployment in real-world scenarios to provide valuable insights into its practical applications and limitations. Additionally, exploring multi-agent systems for the coordination and optimization of multiple shared energy storage operators and distributed energy resources could enhance the overall efficiency and robustness of the grid. Furthermore, studying the impact of consumer behavior and demand response on the effectiveness of low-carbon optimization models could lead to more accurate and user-centric strategies, including analyzing how different incentives and pricing strategies affect consumer participation in demand response programs.

Author Contributions: Conceptualization, L.Z., T.Y., Y.Y., F.P., L.F., S.Q. and Y.H.; methodology, L.Z., T.Y., Y.Y., L.F. and Y.H.; software, L.Z., T.Y., F.P. and S.Q.; validation, L.Z., T.Y., Y.Y., F.P., L.F., S.Q. and Y.H.; formal analysis, L.Z., L.F.; investigation, L.Z.; resources, L.Z.; data curation, L.Z., Y.Y.; writing—L.Z., T.Y.; writing—review and editing, L.Z., T.Y., Y.Y., F.P. and Y.H.; visualization, L.Z., L.F. and T.Y.; supervision, Y.Y., F.P.; project administration, Y.Y., F.P. and Y.H.; funding acquisition Y.Y., F.P. and Y.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Key R&D Program of China (No. 2022YFF0606600), the Science and Technology Project of China Southern Power Grid (GDKJXM20230256), the Energy Bureau of Guangdong Development and Reform Commission (No. 1689833186433), and the Guangdong Basic and Applied Basic Research Foundation (No. 2023A1515012372).

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Conflicts of Interest: Authors Lihua Zhong, Yuyao Yang, Feng Pan, Lei Feng, Shuzhe Qi were employed by the company Metrology Center of Guangdong Power Grid Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. The Metrology Center of Guangdong Power Grid Co., Ltd. had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Yolcan, O.O.J.I.; Development, G. World energy outlook and state of renewable energy: 10-Year evaluation. *Innov. Green Dev.* **2023**, *2*, 100070. [CrossRef]
2. UNFCCC. *Paris Agreement*; UNFCCC: Paris, France, 2015.
3. Kang, C.; Zhou, T.; Chen, Q.; Wang, J.; Sun, Y.; Xia, Q.; Yan, H. Carbon emission flow from generation to demand: A network-based model. *IEEE Trans. Smart Grid* **2015**, *6*, 2386–2394. [CrossRef]
4. Zhang, M.; Xu, Y.; Yi, Z. Two-stage Carbon-Oriented Scheduling of an Active Distribution Network with Thermostatically Controlled Load Aggregators. *IEEE Trans. Sustain. Energy* **2024**, *15*, 1462–1474. [CrossRef]
5. Chen, L.; Zhou, Y. Low carbon economic scheduling of residential distribution network based on multi-dimensional network integration. *Energy Rep.* **2023**, *9*, 438–448. [CrossRef]
6. Zhu, G.; Gao, Y. Multi-objective optimal scheduling of an integrated energy system under the multi-time scale ladder-type carbon trading mechanism. *J. Clean. Prod.* **2023**, *417*, 137922. [CrossRef]
7. Shen, S.; He, D.; Ma, Y.; Suo, X. Sequential Production Simulation Calculation Method of Multi-energy Power System Considering Grid Constraints and Stepped Carbon Emission Trading. In Proceedings of the 2023 10th International Forum on Electrical Engineering and Automation (IFEEA), Piscataway, NJ, USA, 3–5 November 2023; pp. 893–898.
8. Khalid, M. Smart grids and renewable energy systems: Perspectives and grid integration challenges. *Energy Strat. Rev.* **2024**, *51*, 101299. [CrossRef]
9. Xu, B.; Chen, Y.; Shen, X.J.E.R.J. Clean energy development, carbon dioxide emission reduction and regional economic growth. *Econ. Res. J.* **2019**, *54*, 188–202.
10. Sharma, K.K.; Monga, H. Smart Grid: Future of Electrical Transmission and Distribution. *Int. J. Commun. Netw. Syst. Sci.* **2020**, *13*, 45. [CrossRef]
11. Tan, Q.; Ding, Y.; Ye, Q.; Mei, S.; Zhang, Y.; Wei, Y. Optimization and evaluation of a dispatch model for an integrated wind-photovoltaic-thermal power system based on dynamic carbon emissions trading. *Appl. Energy* **2019**, *253*, 113598. [CrossRef]
12. Byrne, R.H.; Nguyen, T.A.; Copp, D.A.; Chalamala, B.R.; Gyuk, I. Energy management and optimization methods for grid energy storage systems. *IEEE Access* **2017**, *6*, 13231–13260. [CrossRef]
13. Eyer, J.; Corey, G. Energy storage for the electricity grid: Benefits and market potential assessment guide. *Sandia Natl. Lab.* **2010**, *20*, 5.
14. Yang, M.; Zhang, Y.; Liu, J.; Yin, S.; Chen, X.; She, L.; Fu, Z.; Liu, H. Distributed Shared Energy Storage Double-Layer Optimal Configuration for Source-Grid Co-Optimization. *Processes* **2023**, *11*, 2194. [CrossRef]
15. Zhu, J.; Li, S.; Borghetti, A.; Lan, J.; Li, H.; Guo, T. Review of demand-side energy sharing and collective self-consumption schemes in future power systems. *iEnergy* **2023**, *2*, 119–132. [CrossRef]
16. Lai, S.; Qiu, J.; Tao, Y. Individualized pricing of energy storage sharing based on discount sensitivity. *IEEE Trans. Ind. Inform.* **2021**, *18*, 4642–4653. [CrossRef]
17. Yan, D.; Chen, Y. Distributed coordination of charging stations with shared energy storage in a distribution network. *IEEE Trans. Smart Grid* **2023**, *99*, 4666–4682. [CrossRef]
18. Du, P.; Huang, B.; Liu, Z.; Yang, C.; Sun, Q. Real-Time Energy Management for Net-Zero Power Systems Based on Shared Energy Storage. *J. Mod. Power Syst. Clean Energy* **2024**, *12*, 371–380. [CrossRef]
19. Xv, A.; He, C.; Zhang, M.; Wang, T. Day-ahead scheduling with renewable generation considering shared energy storage. In Proceedings of the 2022 4th Asia Energy and Electrical Engineering Symposium (AEEES), Chengdu, China, 25–28 March 2022; pp. 492–497.
20. Li, J.; Fang, Z.; Wang, Q.; Zhang, M.; Li, Y.; Zhang, W. Optimal Operation with Dynamic Partitioning Strategy for Centralized Shared Energy Storage Station with Integration of Large-Scale Renewable Energy. *J. Mod. Power Syst. Clean Energy* **2024**, *12*, 359–370. [CrossRef]
21. Miao, A.; Yuan, Y.; Feng, C.; Hou, Y.; Wang, A.; Huang, Y. Low-carbon Economic Scheduling of Park Integrated Energy System Considering User-side Shared Energy Storage. In Proceedings of the 2023 7th International Conference on Smart Grid and Smart Cities (ICSGSC), Lanzhou, China, 22–24 September 2023; pp. 567–572.
22. Shi, Y.; Xu, B.; Wang, D.; Zhang, B. Using battery storage for peak shaving and frequency regulation: Joint optimization for superlinear gains. *IEEE Trans. Power Syst.* **2017**, *33*, 2882–2894. [CrossRef]
23. Wan, T.; Tao, Y.; Qiu, J.; Lai, S. Distributed Energy and Carbon Emission Right Trading in Local Energy Systems Considering the Emission Obligation on Demand Side. *IEEE Syst. J.* **2023**, *99*, 1–10. [CrossRef]
24. Hu, J.; Wang, Y.; Dong, L. Low carbon-oriented planning of shared energy storage station for multiple integrated energy systems considering energy-carbon flow and carbon emission reduction. *Energy* **2024**, *290*, 130139. [CrossRef]
25. Zeng, L.; Gong, Y.; Xiao, H.; Chen, T.; Gao, W.; Liang, J.; Peng, S. Research on interval optimization of power system considering shared energy storage and demand response. *J. Energy Storage* **2024**, *86*, 111273. [CrossRef]
26. Hao, X.; Chen, Y.; Wang, H.; Wang, H.; Meng, Y.; Gu, Q. A V2G-oriented reinforcement learning framework and empirical study for heterogeneous electric vehicle charging management. *Sustain. Cities Soc.* **2023**, *89*, 104345. [CrossRef]
27. Wang, Y.; Cui, Y.; Li, Y.; Xu, Y. Collaborative optimization of multi-microgrids system with shared energy storage based on multi-agent stochastic game and reinforcement learning. *Energy* **2023**, *280*, 128182. [CrossRef]
28. Zhang, J.; Sang, L.; Xu, Y.; Sun, H. Networked Multiagent-Based Safe Reinforcement Learning for Low-Carbon Demand Management in Distribution Networks. *IEEE Trans. Sustain. Energy* **2024**, *15*, 1528–1545. [CrossRef]

29. Wang, H.; Wang, Y.; Lu, Y.; Fu, Q.; Chen, J. Visual interpretation of deep deterministic policy gradient models for energy consumption prediction. *J. Build. Eng.* **2023**, *79*, 107847. [CrossRef]
30. Lillicrap, T.P.; Hunt, J.J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; Wierstra, D. Continuous control with deep reinforcement learning. *arXiv* **2015**, arXiv:1509.02971.
31. Kang, D.; Kang, D.; Hwangbo, S.; Niaz, H.; Lee, W.B.; Liu, J.J.; Na, J. Optimal planning of hybrid energy storage systems using curtailed renewable energy through deep reinforcement learning. *Energy* **2023**, *284*, 128623. [CrossRef]
32. Yang, Y.; Li, M.; Lu, Q.; Xie, P.; Wei, W. Real-Time Operation of Energy Storage Assisting Utilization of Offshore Wind Power: Improving the Lyapunov Policy via DDPG. In Proceedings of the 2022 12th International Conference on Power and Energy Systems (ICPES), Guangzhou, China, 23–25 December 2022; pp. 731–736.
33. Han, X.; Shi, W.; Yuan, X.; Hao, Y. Multi-scene Scheduling of Power System with Renewable Energy Based on DDPG. In Proceedings of the 2023 8th Asia Conference on Power and Electrical Engineering (ACPEE), Tianjin, China, 14–16 April 2023; pp. 1892–1897.
34. Li, Y.; Wu, J.; Pan, Y. Deep Reinforcement Learning for Online Scheduling of Photovoltaic Systems with Battery Energy Storage Systems. *Intell. Converg. Networks* **2024**, *5*, 28–41. [CrossRef]
35. Bhatti, H.J.; Danilovic, M.J.W.J.o.E.; Technology. Making the world more sustainable: Enabling localized energy generation and distribution on decentralized smart grid systems. *J. Eng. Technol.* **2018**, *6*, 350–382. [CrossRef]
36. Zhang, Y.; Han, Y.; Liu, D.; Dong, X. Low-carbon Economic Dispatch of Electricity-Heat-Gas Integrated Energy Systems Based on Deep Reinforcement Learning. *J. Mod. Power Syst. Clean Energy* **2023**, *11*, 1827–1841. [CrossRef]
37. Luo, Y.; Hao, H.; Yang, D.; Zhou, B. Multi-objective Optimization of Integrated Energy Systems Considering Ladder-type Carbon Emission Trading and Refined Load Demand Response. *J. Mod. Power Syst. Clean Energy* **2024**, *12*, 828–839. [CrossRef]
38. Baxter, L.A.; Puterman, M.L. Markov Decision Processes: Discrete Stochastic Dynamic Programming. *Technometrics* **1995**, *37*, 353. [CrossRef]
39. Schaul, T.; Quan, J.; Antonoglou, I.; Silver, D. Prioritized Experience Replay. *arXiv* **2015**, arXiv:1511.05952.
40. Bansal, N.; Chen, X.; Wang, Z. Can We Gain More from Orthogonality Regularizations in Training Deep CNNs? *arXiv* **2018**, arXiv:1810.09102.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.