Electronic Research Archive

https://www.aimspress.com/journal/era

ERA, 34(6): 3768–3789.
DOI: 10.3934/era.2026170
Received: 06 February 2026
Revised: 11 April 2026
Accepted: 22 April 2026
Published: 06 May 2026

Research article

A two-level load-utility balancing model for multi-source smart grids based on multi-agent reinforcement learning algorithm

Linsen Song[1,*], **Yukai Zhang**[1] **and Jishen Jia**[2]

[1] School of Artificial Intelligence, Henan Institute of Science and Technology, Xinxiang 453003, China
[2] School of Mathematical Sciences, Henan Institute of Science and Technology, Xinxiang 453003, China

* **Correspondence:** Email: slinsen@163.com.

Abstract: In smart grids, the interactions among diverse participants significantly affect system efficiency and social welfare. Considering that the real-time electricity pricing (RTP) mechanism under traditional social welfare maximization fails to account for the independent interests of various entities in the power system, a two-level load-utility balancing model for real-time pricing in smart grids is proposed in this paper, where the welfare of the demand side and the multi-energy supply side is collaboratively optimized and balanced. Furthermore, a multi-agent reinforcement learning (MARL) algorithm based on the centralized training and decentralized execution (CTDE) framework is designed for this model, and a multi-agent electricity market environment is constructed accordingly, comprising users, an aggregated power supplier, and a power market scheduling center (PMSC). The user agent is modeled with a heterogeneous utility function, the supplier agent is modeled with a profit function coordinating both traditional and renewable energy, while the PMSC agent is responsible for real-time pricing and cross-agent welfare balance coordination. Finally, simulation results show the effectiveness of the proposed model and algorithm in achieving welfare balance between the supplier and users. Compared with the pricing scheme without a welfare-balancing mechanism, the proposed model reduces the welfare gap between the supplier and users by approximately 46.9%. Compared with the non-dominated sorting genetic algorithm II (NSGA-II), the proposed method can achieve a comparable level of total social welfare.

Keywords: smart grid; real-time pricing; social welfare; reinforcement learning

1. Introduction

In recent years, the excessive use of fossil fuels has led to a series of global challenges, including resource depletion, climate change, and geopolitical conflicts, driving energy systems toward low-carbon and sustainable transitions. Motivated by carbon reduction targets, energy security concerns, and tech-

nological advancements, renewable energy sources—particularly wind and photovoltaic power—have experienced rapid growth in both installed capacity and penetration within power systems. However, due to the intermittency and stochastic nature of renewable generation, their large-scale integration, coupled with increasingly stringent environmental constraints, poses more complex challenges to real-time power balancing and coordinated system operation. To cope with these challenges, integrated energy systems that combine traditional and renewable energy in a complementary manner are regarded as an important solution to enhance flexibility and reliability. This process also relies on smart grids, which achieve efficient and reliable electricity supply through system-wide intelligent control. Within smart grids, demand-side management (DSM) is considered a key approach to improving system flexibility, as it enables users to adjust their electricity consumption in response to grid conditions. Among various DSM strategies, price-based demand response mechanisms play a central role. Common pricing schemes include time-of-use pricing, critical peak pricing, and RTP. While the first two can achieve peak shaving and valley filling to some extent, their flexibility is limited. In contrast, RTP dynamically reflects real-time system conditions and allows more effective coordination among multiple stakeholders, which explains why it has attracted increasing attention in recent years. Despite the growing interest in RTP-based demand-side management, many existing studies adopt simplified assumptions on the supply side, typically modeling electricity providers as relying on a single type of energy source. Such assumptions become increasingly restrictive as renewable energy penetration rises, since the fundamental trade-off between the stability of traditional generation and the uncertainty of renewable energy cannot be adequately captured within a single-source framework. As a result, these models may fail to fully reflect the operational flexibility, pricing behavior, and welfare implications of modern power systems.

With the increasing share of renewable energy and the growing demand for flexible load management, electricity pricing has attracted significant attention. Existing studies on RTP mainly follow two approaches: game-theoretic models and optimization-based methods. Game-theoretic studies analyze strategic interactions between supply and demand by formulating multi-level games and deriving equilibrium outcomes, such as Nash equilibria [1–4]. However, these equilibria are not necessarily socially optimal and may lead to efficiency losses. Optimization-based approaches, on the other hand, directly formulate RTP as a social welfare maximization problem [5–10]. Among them, bilevel optimization models the hierarchical decision-making process between system operators and market participants under the RTP framework. While such formulations can accurately capture multi-level interactions, they often suffer from high computational complexity, convergence difficulties, and conflicting objectives between different decision levels. For example, Wang et al. [8] developed a demand response–based real-time pricing approach that formulates social welfare maximization as a Karush-Kuhn-Tucker (KKT) complementarity problem and solves it using a Jacobian smoothing Newton method, enabling coordinated optimization of electricity price, consumption, and production in smart grids. Li et al. [9] proposed a real-time electricity pricing solution method based on smoothing approximation and Newton iteration for solving nonlinear multi-user real-time pricing models in smart grids. Song et al. [10] proposed a distributed optimization method for multi-source power supply and demonstrated, through numerical studies, the advantages of multi-source models over single-source formulations in terms of system efficiency, providing useful insights for scheduling and pricing in smart grids. The second-order cone algorithm proposed by Xu et al. [11] effectively solves the non-convex and nonlinear problems in power flow optimization of distribution networks, and provides a mathematically rigorous modeling framework for core physical constraints such as node voltage stability and branch

transmission capacity limits in active distribution networks. Although social welfare maximization avoids efficiency losses caused by conflicting individual objectives and aligns well with regulatory goals of electricity markets, these optimization-based methods typically rely on accurate prior information and exhibit limited scalability and adaptability under large-scale uncertainty. Moreover, maximizing total welfare may inadvertently compromise the welfare of certain market participants. Therefore, combining social welfare optimization with mechanisms that can flexibly balance individual objectives under uncertainty remains a challenging and important research problem in electricity market design.

Model-free reinforcement learning algorithms do not require an explicit mathematical model of the system. Instead, policies are continuously updated through interactions with dynamic environments, enabling agents to directly learn optimal or near-optimal decision strategies under uncertainty [12, 13]. Owing to its learning-based nature, several studies have proposed applying reinforcement learning (RL) techniques to pricing and energy management [14–19]. Lu et al. [14] were the first to apply reinforcement learning to RTP problems. Wang et al. [15] proposed a reinforcement learning-based bilevel real-time pricing strategy that effectively balances the interests of power suppliers and consumers. Song et al. [16] used MARL to address the differences in load characteristics and utility preferences among diverse electricity users in smart grids, balancing the interests of power suppliers and users while ensuring welfare equity across different users. Ahrarinouri et al. [17] proposed a MARL approach for residential energy consumption management. Lai et al. [18] developed a MARL-based community energy management system to reduce peak increases caused by renewable energy uncertainty. He et al. [19] designed a prioritized experience replay-based deep deterministic policy gradient (Per-DDPG) algorithm to maximize supply–demand balance and carbon reduction benefits at the same time. Du et al. [20] adopted the multi-agent deep deterministic policy gradient algorithm to solve the day-ahead electricity market bidding problem, and approximated the Nash equilibrium through a centralized training and decentralized execution mechanism. By modeling the pricing problem as a sequential decision process, RL can adaptively learn optimal strategies through interaction with the environment. Compared with traditional methods, RL-based approaches rely less on prior assumptions, adapt better to uncertainty, and show great potential in promoting renewable energy integration and maximizing social welfare. Deep reinforcement learning (DRL) uses artificial neural networks (ANNs) as function approximators, enabling agents to learn continuous state–action transitions under uncertainty and extract hidden features from high-dimensional state spaces [21]. The convergence performance and generalization ability of DRL models are highly dependent on the rationality of hyperparameter settings, and existing studies have carried out in-depth exploration on hyperparameter optimization methods for machine learning models [22].

Although recent studies have explored RL for electricity price optimization, peak shaving, and demand response, the application of MARL to RTP with explicit modeling of multi-party interactions remains limited. In practice, the power system is not a single decision-making entity but a complex interactive system composed of multiple participants, including user agents, supplier agents, and system-level dispatchers. The strategies of these agents are mutually coupled, and system-level performance depends on coordinated decision-making rather than the optimization of any individual agent. Driven by increasing energy demand and advances in smart grid technologies, power systems are evolving from traditional one-way transmission structures toward highly interactive, multi-agent frameworks, which poses new challenges to conventional pricing and optimization methods. To address these challenges, this paper adopts a CTDE MARL framework for RTP. This framework is highly suitable for the tripartite

decision-making characteristics of the electricity market, which not only realizes independent local decision-making of each agent with effective privacy protection, but also achieves global coordinated optimization via centralized training. It explicitly captures strategic interactions among users, suppliers, and the system operator while incorporating system-level objectives such as supply–demand balance and operational stability. Compared with traditional bilevel programming, numerical optimization, or game-theoretic models—whose structures are often tightly coupled to fixed market settings—the proposed MARL-based approach offers superior scalability and flexibility. It does not rely on analytical solvers, allows new market participants or rules to be introduced without rederiving mathematical models, and enables multiple objectives, including supplier cost constraints, user comfort preferences, system balancing requirements, and social welfare maximization, to be directly embedded into the reward design. Through continuous interaction with the environment, agents can autonomously learn coordinated strategies, achieving multi-objective optimization with reduced computational burden and supporting near real-time decision-making in complex and uncertain electricity markets.

To balance user and supplier welfare while maximizing social welfare, different market participants in the power system are modeled as autonomous agents, each pursuing its own objectives. Specifically, three types of agents are designed: the system operator, the supplier, and the users. The system operator, as the core entity responsible for ensuring secure and stable system operation, integrates market information and determines optimal prices to achieve efficient resource allocation. The supplier is responsible for electricity generation. Due to the inherent variability and uncertainty of renewable energy, it must determine the optimal level of traditional power output based on renewable generation forecasts. Users, as demand-side participants, autonomously decide their electricity consumption to maximize their individual welfare. Renewable energy forecasting is very challenging because renewable output depends heavily on the weather and changes unpredictably. If the renewable energy output forecast is not accurate, it will reduce the effectiveness of pricing and dispatch to some extent. Therefore, instead of relying on mathematical assumptions to construct stochastic prediction models as in previous studies, this paper utilizes a reliable renewable generation dataset to ensure the accuracy and robustness of the results [23].

The main contributions of this paper are as follows. 1) A real-time pricing model oriented toward multi-objective welfare balance is proposed, and a tripartite interaction framework is established under the Markov decision process (MDP) paradigm. The model takes into account multiple types of electricity prices, renewable energy integration, and grid constraints, solving the core defects of traditional pricing models that ignore the independent interests of multiple entities and adopt oversimplified single-source supply-side assumptions, providing a novel approach to real-time pricing in smart grids. 2) By accounting for the preference heterogeneity among diverse market participants and introducing a dedicated welfare equilibrium mechanism, the proposed approach effectively avoids unfair welfare distribution caused by solely maximizing total social welfare, achieving a relative welfare balance between residential users and power suppliers while ensuring overall system efficiency. 3) To address the limitation that existing MARL-based RTP studies mostly adopt oversimplified two-party game modeling, we model each market participant as an independent decision-making agent, customize exclusive state space, action space, and reward function for each agent, and build a tripartite interactive electricity market simulation environment to realize coordinated optimal decision-making via the CTDE-based MARL framework. 4) The differences between renewable and traditional energy sources in terms of carbon emission characteristics are considered, and the carbon emission cost is incorporated into the

generation cost, with distinct utility parameters and usage preferences designed for different users and different types of energy.

The structure of this paper is as follows: Section 2 presents the mathematical formulation of the system model; Section 3 reformulates the model as a Markov decision process; in Section 4, we propose a multi-agent reinforcement learning algorithm for RTP; and Section 5 validates the model and algorithm through numerical experiments.

2. System model

Consider a smart grid system composed of an aggregated power supplier, a PMSC, and multiple residential users. The supplier provides both traditional and renewable energy. Traditional energy, generated primarily from fossil fuels such as coal, is controllable and dispatchable. Renewable energy, mainly wind power, can be predicted from weather conditions, but it still cannot be controlled. The PMSC plays a central role, aiming to maximize social welfare while maintaining supply–demand balance. The operation horizon is divided into k discrete time slots. Let the set of users be $N = \{1, 2, ..., n\}$ and the set of time slots be $K = \{1, 2, ..., k\}$. At the beginning of each time slot, all agents make decisions at the same time based on their own observations. The PMSC sets the prices of traditional energy and renewable energy based on its observation. The supplier decides the amount of traditional energy generation based on its observation. Each user decides its demand for both types of energy according to its observation. All actions are sent to the environment in the same step. Then the environment clears the market to match supply and demand and generates the next observations. The interactions among participants in this smart grid system are illustrated in Figure 1.

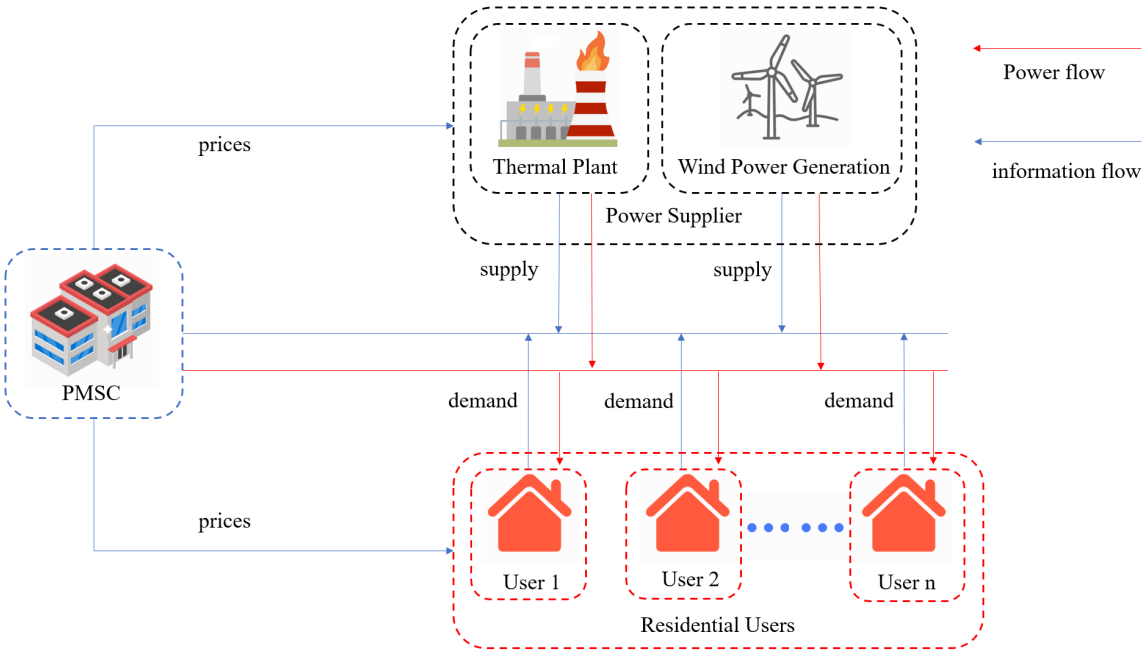


Figure 1. Topology figure of smart grid system.

Consider a smart grid with R residential users, where the dispatcher sets the prices for the two types

of electricity. Let $x_i^{k,T}$ and $x_i^{k,N}$ denote the traditional energy demand and renewable energy demand of user i in the period k, respectively. The total power consumption of residential users i during the period of k is

$$x_i^k = x_i^{k,T} + x_i^{k,N}, \tag{2.1}$$

where $x_{min} \le x_i^k \le x_{max}$, x_{min}, x_{max} are the minimum and maximum power demand, respectively. Let L_k^T be the traditional energy supply during the period k, which satisfies

$$L_{min}^T \le L_k^T \le L_{max}^T. \tag{2.2}$$

Let L_k^N be the new energy supply in the period k; the constraint on the Wind output is

$$L_{min}^N \le L_k^N \le L_{max}^N. \tag{2.3}$$

2.1. Demand side: Welfare model of residential users

Residential users are the key participants on the demand side of the electricity market. The benefits of users consist of the following two parts: The first part is the utility generated from electricity consumption. Different utility functions are introduced for the two types of electricity.

$$U_T(x_i^{k,T}, \omega_i^{k,T}) = \begin{cases} \omega_i^{k,T} x_i^{k,T} - \frac{\alpha_i}{2}(x_i^{k,T})^2, 0 \le x_i^{k,T} \le \frac{\omega_i^{k,T}}{\alpha_i} \\ \frac{(\omega_i^{k,T})^2}{2\alpha_i}, \qquad\qquad x_i^{k,T} \ge \frac{\omega_i^{k,T}}{\alpha_i}, \end{cases} \tag{2.4}$$

$$U_N(x_i^{k,N}, \omega_i^{k,N}) = \begin{cases} \omega_i^{k,N} x_i^{k,N} - \frac{\alpha_i}{2}(x_i^{k,N})^2, 0 \le x_i^{k,N} \le \frac{\omega_i^{k,N}}{\alpha_i} \\ \frac{(\omega_i^{k,N})^2}{2\alpha_i}, \qquad\qquad x_i^{k,N} \ge \frac{\omega_i^{k,N}}{\alpha_i}. \end{cases} \tag{2.5}$$

The utility functions satisfy two properties: non-decreasing and diminishing marginal utility, namely $\frac{\partial U(\cdot)}{\partial x} \ge 0, \frac{\partial U^2(\cdot)}{\partial^2 x} \le 0$. $x_i^{k,T}$ and $x_i^{k,N}$ mean the traditional energy demand and renewable energy demand of user i at time slot k, respectively. ω is the preference level. In electricity demand models, it is often understood as how much the user wants to use electricity. α is commonly regarded as the rate of diminishing marginal utility and, in demand response studies, is frequently interpreted as the user's price sensitivity. To capture the differences in users' willingness to consume electricity over time, the values of ω typically vary across both users and time periods. To reflect the differences in price sensitivity, the value of α_i is user-specific. $\omega_i^{k,T}$ and $\omega_i^{k,N}$ represent the utility parameters of user i for traditional energy and renewable energy in period k, respectively, which satisfy

$$\alpha_{min} \le \alpha_i \le \alpha_{max}, \tag{2.6}$$

$$\omega_{min}^T \le \omega_i^{k,T} \le \omega_{max}^T, \tag{2.7}$$

$$\omega_{min}^N \le \omega_i^{k,N} \le \omega_{max}^N. \tag{2.8}$$

$U_i^k = U_T(x_i^{k,T}, \omega_i^{k,T}) + U_N(x_i^{k,N}, \omega_i^{k,N})$ denotes the total utility of user i during period k.

The second part focuses on the electricity cost of users. Denote p_k^T, p_k^N as the real-time price of the traditional energy power and the renewable energy power, respectively, where each price is constrained by $p_{\min} \le p_k^T, p_k^N \le p_{\max}$. The electricity cost of the residential user i in the period k is:

$$C_i^k = p_k^T \cdot x_i^{k,T} + p_k^N \cdot x_i^{k,N}. \tag{2.9}$$

In summary, the total welfare of the residential users in period k is:

$$W_k^R = \sum_{i=1}^{n} (U_i^k - C_i^k). \tag{2.10}$$

2.2. Supply side: Welfare model of power supplier

Clean energy has become an essential component of modern power grids. However, its output is inherently uncertain due to its strong dependence on weather conditions, making it difficult for renewable sources alone to guarantee stable system operation. In contrast, traditional energy sources offer stable and controllable generation but produce substantial carbon emissions. Under the current "dual-carbon" policy goals, the use of such high-emission energy sources is increasingly restricted and regulated [24] . Therefore, an aggregated supplier that integrates both renewable and traditional energy sources can achieve higher overall benefits by balancing low-carbon objectives with system reliability. This aggregated supplier provides both types of electricity to users and earns revenue through electricity sales. Its total revenue consists of the following four components.

The first part is the revenue obtained by the supplier through electricity sales, which is equal to the total electricity purchase cost of all users. The supplier's sales revenue at time slot k is given by:

$$B_k = \sum_{i=1}^{n} (p_k^T \cdot x_i^{k,T} + p_k^N \cdot x_i^{k,N}). \tag{2.11}$$

The second part is the generation cost of electricity from traditional energy sources. Since the cost increases with the amount of electricity generated, the cost of traditional power generation is modeled by the following function:

$$C_1^k(L_k^T) = a(L_k^T)^2 + bL_k^T + c, \tag{2.12}$$

where L_k^T denotes the total amount of traditional electricity generated in time slot k. The coefficients $a > 0$, $b \geq 0$, and $c \geq 0$ are known parameters.

The third part is the carbon tax that the power supplier needs to pay for producing L_k^T traditional energy power. The actual carbon emission of L_k^T is $Q^c = r(L_k^T)^2 + sL_k^T + t$. This quadratic carbon emission function is constructed based on the emission characteristics of traditional thermal power generation. Let the carbon tax price per ton of carbon dioxide be p^c, and then the carbon tax that the power supplier needs to pay is

$$C_2^k(L_k^T) = p^c \cdot Q^c = p^c \cdot [r(L_k^T)^2 + sL_k^T + t]. \tag{2.13}$$

The forth part is the generation cost of electricity from renewable energy sources. As renewable resources are naturally available, their marginal generation cost can be considered near-zero. Thus, only fixed costs and equipment maintenance costs are taken into account. Accordingly, the renewable power generation cost function is modeled as the following function:

$$C_3^k(L_k^N) = \theta(L_k^N)^2 + \eta L_k^N. \tag{2.14}$$

In summary, the total welfare of the power supplier in period k can be formulated as follows:

$$W_k^S = B_k - C_1^k(L_k^T) - C_2^k(L_k^T) - C_3^k(L_k^N). \tag{2.15}$$

3. MDP formulation

3.1. Definition of state, action, and reward in MDP

To implement the model, we use reinforcement learning algorithms to simulate the interactions among the three participants in the electricity market. For reinforcement learning implementation, the model is reformulated as an MDP. Within this framework, three types of market participants—residential users, the aggregated power supplier, and the power market scheduling center (PMSC)—are each modeled as an independent agent that learns its optimal strategy through interactions with the environment. Specifically, the MDP consists of four fundamental elements: the set of all possible states S, the set of all possible actions A, the set of rewards R obtained by performing a specific action in a given state, and the state transition probabilities P [25].

3.2. Transforming the demand-side model into an MDP

Residential users determine their own load based on the current electricity price and the utility generated by the electricity. The state of residential users is defined by the following elements: the time slot t, the traditional electricity price, the renewable electricity price, the historical generation of traditional energy, the historical demand for traditional energy, the historical generation of renewable energy, and the historical demand for renewable energy. To capture the periodicity of time (and avoid discontinuities such as 23:00 vs 00:00), we encode time t as $t_{sin} = sin(2\pi t/T)$ and $t_{cos} = cos(2\pi t/T)$, where T is the period (e.g., 24 for hour-of-day). This cyclical encoding is standard in time-series feature engineering and positional encoding literature. The specific expression is given as follows:

$$\begin{aligned} S_{user} = [&t_{sin}, t_{cos} \\ &traditional\ electricity\ price, \\ &renewable\ electricity\ price, \\ &historical\ supply\ of\ traditional\ energy, \\ &historical\ supply\ of\ renewable\ energy, \\ &historical\ demand\ of\ traditional\ energy, \\ &historical\ demand\ of\ renewable\ energy]. \end{aligned} \tag{3.1}$$

In a given state, the action taken by the residential user at time k is to determine its traditional and renewable electricity demand, which can be formally expressed as

$$\begin{aligned} a_1^{k,R} &= \sum_{i=1}^{n} x_i^{k,T}, \\ a_2^{k,R} &= \sum_{i=1}^{n} x_i^{k,N}, \end{aligned} \tag{3.2}$$

where $a_1^{k,R}$ denotes the first action of the residential user at time period k, and $a_2^{k,R}$ denotes the second action of the residential user at time period k, $x_i^{k,T}$ represents the demand of residential user i for traditional electricity at time k, and $x_i^{k,N}$ represents the demand of residential user i for renewable electricity at time k. Based on this, the reward for the residential user is

$$R_k^R = \alpha^R W_k^R, \tag{3.3}$$

where α is a scaling coefficient. To ensure stable training and prevent domination by agents with larger reward values, the rewards of different agents are scaled to a comparable range. Specifically, each agent's original reward is multiplied by a predefined scaling coefficient. This scaling does not change the optimization objective but improves numerical stability during learning.

3.3. Transforming the supply-side model into an MDP

The supplier determines its traditional energy supply based on the electricity price, generation cost, wind power output, and other relevant information. While determining its traditional power generation, the supplier submits a reference minimum electricity price to the dispatching center based on its own generation cost, ensuring that its minimum profit is secured. We define the supplier's state as follows: cyclic time features t_{sin} and t_{cos}, the price of traditional electricity, the historical supply of traditional energy, the current supply of renewable energy, and the historical trading volume. The specific expression is given as follows:

$$\begin{aligned} S_{supplier} = [&t_{sin}, t_{cos} \\ &traditional\ electricity\ price, \\ &historical\ supply\ of\ traditional\ energy, \\ &current\ supply\ of\ renewable\ energy, \\ &historical\ trading\ volume]. \end{aligned} \tag{3.4}$$

In a given state, the action taken by the supplier at time k is to generate traditional electricity, which can be formally represented as

$$a^{k,S} = L_k^T. \tag{3.5}$$

In a given state, the reward obtained by taking the corresponding action can be expressed as

$$R_k^S = \alpha^S W_k^S. \tag{3.6}$$

3.4. Transforming the PMSC model into an MDP

The primary responsibility of the PMSC is to ensure the stable operation of the electricity market while enhancing overall social welfare. Accordingly, social welfare and supply–demand balance are defined as the objectives of the PMSC, which can be regarded as its own reward. The PMSC sets the prices of traditional and renewable electricity at the current time based on historical residential user demand, supplier generation, historical prices, the minimum price of traditional energy, and the minimum price of renewable energy provided by the supplier and the forecasted renewable generation. These prices are then communicated to the residential users and the supplier. The current state of PMSC is characterized by the following elements: cyclic time features t_{sin} and t_{cos} [26], the traditional electricity price, the renewable electricity price, the historical generation of traditional energy, the historical demand for traditional energy, the historical generation of renewable energy, the historical demand for renewable energy, the minimum price of traditional energy, and the minimum price of

renewable energy. The specific expression is given as follows:

$$\begin{aligned} S_{PMSC} = [&t_{sin}, t_{cos} \\ &traditional\ electricity\ price, \\ &renewable\ electricity\ price, \\ &historical\ supply\ of\ traditional\ energy, \\ &historical\ supply\ of\ renewable\ energy, \\ &historical\ demand\ of\ traditional\ energy, \\ &historical\ demand\ of\ renewable\ energy, \\ &The\ minimum\ price\ of\ traditional\ energy, \\ &The\ minimum\ price\ of\ renewable\ energy]. \end{aligned} \tag{3.7}$$

In a given state, the actions taken by the PMSC at time k are to set the price of traditional electricity and the price of renewable electricity, which can be formally expressed as

$$a_1^{k,P} = p_k^T, \tag{3.8}$$

$$a_2^{k,P} = p_k^N. \tag{3.9}$$

The reward of the PMSC consists of two main components and several penalty terms. The first component is the social welfare term, which aims to maximize the overall social welfare while reducing the welfare difference between suppliers and users. The second component is the supply–demand balance term, which penalizes the mismatch between supply and demand for both traditional and renewable energy. A slightly larger weight is assigned to the renewable energy mismatch to ensure its priority consumption. In addition, several penalty terms are incorporated to enforce practical constraints, including a penalty on electricity price fluctuations to promote price stability, as well as penalties imposed when the welfare of residential users or suppliers falls below predefined lower bounds, thereby guaranteeing minimum welfare levels for both parties. In a given state, the reward obtained by taking the corresponding action can be expressed as follows:

$$R_1^P = W_k^R + W_k^S - \left|W_k^R - W_k^S\right|, \tag{3.10}$$

$$R_2^P = \left|\sum_{i=1}^{n} x_i^{k,N} - L_k^N\right|^2 + \left|\sum_{i=1}^{n} x_i^{k,T} - L_k^T\right|^2, \tag{3.11}$$

$$R_k^P = \alpha^P R_1^P + \beta^P R_2^P, \tag{3.12}$$

where α^P and β^P are weighting parameters. Among them, β^P is negative, and R_2^P is a deduction term in the reward function. Actually, there are some other penalty terms here, which are specifically designed to avoid extreme welfare values and unstable training dynamics, such as the price fluctuation penalty, the renewable energy pricing penalty, and the negative welfare penalty. The price fluctuation penalty is designed to avoid sharp fluctuations in traditional and renewable energy prices between adjacent time slots and improve market predictability, which is defined as the sum of the absolute differences of the two energy types between the current and previous time slot with the form $\left|P_k^T - P_{k-1}^T\right| + \left|P_k^N - P_{k-1}^N\right|$.

The renewable energy pricing penalty includes two parts: a penalty of 10 times the price difference if $p_k^N > p_k^T$, and an additional fixed penalty of 10 if renewable energy output is below the preset threshold and $p_k^N < 1$. The negative welfare penalty is calculated as the sum of the absolute values of negative welfare of residential users or the negative profit of the power supplier.

4. Algorithm design

RL addresses the problem of how an agent can interact with a dynamic and uncertain environment to learn a policy that maximizes long-term cumulative rewards. Specifically, at each time step, the agent selects an action based on the current state of the environment, which then returns a new state along with a corresponding reward signal. The agent continuously optimizes its policy accordingly, thereby improving decision-making performance and maximizing the expected return. This approach possesses the capabilities of self-learning, self-adaptation, and continuous improvement. In a multi-agent system (MAS), multiple agents interact with the environment sequentially. In this setting, after each agent takes an action, the environment updates its state and returns the corresponding reward signal, while the behavior of preceding agents influences the observations and decisions of subsequent agents. Each agent still follows its original objective but must continuously adjust its strategy under the dynamic environment and the effects of other agents' prior actions, thereby forming a more complex overall operating mechanism. The interaction framework for the electricity market is illustrated in Figure 2.

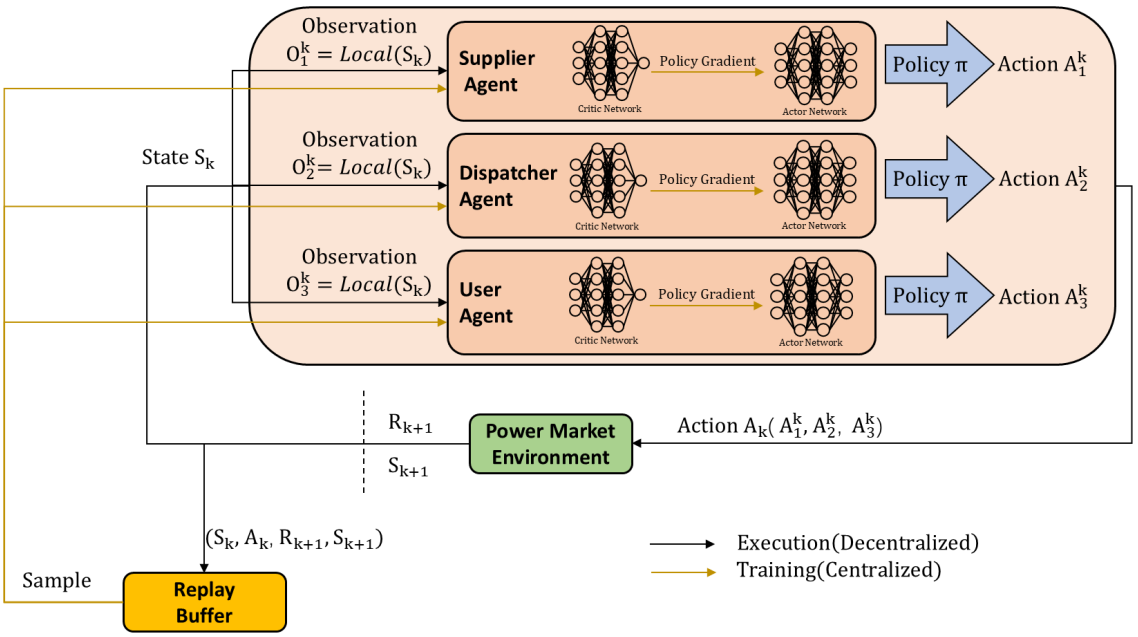


Figure 2. Multi-agent interaction framework for the electricity market.

Figure 2 shows the overall framework of the proposed MARL model for the electricity market. The electricity market is modeled as the environment, which represents the operation and clearing process of the power system. Three different agents interact with this environment: a supplier agent, a

dispatcher agent, and a user agent. Each agent represents one type of market participant and makes its own decisions independently. At each time step k, each agent receives its own observation O_i^k, which contains only the information available to that agent. Based on its local observation, each agent uses an actor network to generate an action according to its policy. The action of each agent is denoted by A_i^k. All agents make their decisions at the same time, and their actions are combined into a joint action vector $A_k = (A_1^k, A_2^k, A_3^k)$. This joint action is then sent to the environment. After receiving the joint action, the environment executes market rules. $S_k = (S_{user}, S_{supplier}, S_{PMSC})$ represents the global state, which aggregates the local observed states of all three agents into a complete system-wide information set. The environment then executes the state transition from the global state S_k to the next global state S_{k+1}, and outputs the corresponding individual reward signals R_{k+1}. The transition $(S_k, A_k, R_{k+1}, S_{k+1})$ is stored in the replay buffer for training, to break time-series data correlation, reuse historical experience, and stabilize multi-agent joint training convergence. During training, each agent has a critic network. Unlike the actor, the critic can access the global state and the joint actions of all agents. The critic evaluates how good the current actions are in terms of long-term rewards. Based on this evaluation, the critic provides policy gradient information to update the actor network. Through this process, the actor gradually learns better decision policies. During actual operation, only the actor networks are used to make decisions, and no global information is required. This training method follows the CTDE principle, which allows effective learning while keeping the decision process practical and scalable. This principle ensures that the MARL model proposed in this study achieves distributed decision-making with privacy protection. During the execution phase, each agent makes decisions solely based on its own local observations, without disclosing any private information to other agents or a centralized scheduling node throughout the entire process.

The CTDE framework is adopted to alleviate the endogenous nonstationarity and convergence fluctuations caused by the concurrent strategy updates of multiple agents. The theoretical rationality and effectiveness of this framework have been extensively demonstrated in the classical multi-agent deep deterministic policy gradient (MADDPG) algorithm [27]. In multi-agent scenarios, the root cause of non-stationarity lies in the continuous evolution of each agent's policy π_i during training, which results in a dynamically changing state transition probability $P(s' \mid s, a_i)$ from the perspective of an individual agent. To address this, our algorithm introduces the joint action of all agents $\mathbf{a} = \{a_1, \ldots, a_N\}$ as input to the critic. This design leverages an inherent property of environment dynamics: The state transition probability satisfies $P(s' \mid s, a_1, \ldots, a_N, \pi_1, \ldots, \pi_N) = P(s' \mid s, a_1, \ldots, a_N)$. That is, given the joint action, the environment follows fixed Markovian transition dynamics independent of individual policy updates. This fundamentally mitigates the core barrier of "environmental nonstationarity caused by policy updates" from a mathematical perspective, and significantly reduces the bias and variance of gradient estimation compared to decentralized training. While a complete theoretical proof of convergence in general nonstationary MARL remains an open problem, our CTDE-based design ensures that from each agent's perspective, the environment appears stationary conditioned on the joint action, allowing standard policy gradient convergence results to apply.

On this basis, we further stabilize the strategy update process via three mechanisms: (i) different learning rates for actor and critic; (ii) a fixed soft update coefficient and an experience replay buffer to break temporal correlations and reduce update variance; and (iii) iterative hyperparameter calibration based on MADDPG benchmarks, with convergence stability and optimization effect as the core objectives. As illustrated in Figure 3, the total system reward converges to a stable plateau after approximately

4000 training episodes and remains stable during 1000 validation episodes with frozen network weights, directly demonstrating the convergence and steady-state stability of our algorithm.

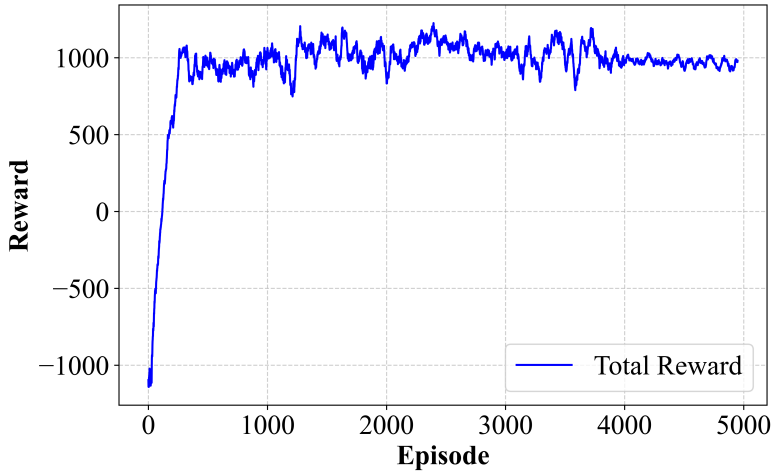


Figure 3. Convergence curve of the system reward during MARL training.

5. Numerical simulation

In this section, we present a detailed analysis of the simulation results of the RTP MARL model. The simulations were conducted on a computer system equipped with an Intel Core i5-9300H processor and 16 GB of memory.

The proposed model consists of one supplier, one grid dispatcher, and 50 heterogeneous residential users, with the supplier providing two types of electricity: traditional thermal power and wind power. The model runs in 24-hour daily cycles, with a total of 5000 iterations: The first 4000 iterations are for model training, and the remaining 1000 iterations are for validation with no neural network updates. On the demand side, residential users are assigned distinct utility function parameters to reflect heterogeneity: The preference parameters for traditional and wind energy vary across both users and time periods (off-peak, flat, and peak hours), while the marginal utility diminishing coefficient only differs across users and remains time-invariant. On the supply side, the supplier aims to maximize its own profit, with wind power output data derived from a real-world wind farm dataset with an added disturbance term. As the core regulator of the system, the dispatcher is designed to maximize overall social welfare, while ensuring the stable operation of the smart grid, maintaining supply-demand balance, and accounting for the benefits of all market participants.

For the adopted MADDPG algorithm, we set a higher learning rate for the critic network to enable fast adaptation to the dynamic environment and a lower learning rate for the actor network to ensure stable policy updates. The action noise is set to decay gradually during training to balance exploration and exploitation in the continuous action space. During validation, the action noise is disabled to fix the policy for accurate performance evaluation. Detailed parameter settings of the system model and training process are summarized in Tables 1 and 2, respectively. Notably, all agents in the MADDPG framework adopt an identical network architecture with consistent structural parameters.

Table 1. Key parameter settings of the system model.

Notations	Value	Description
$[x_{min}, x_{max}]$	[1, 4] kW	Power consumption bounds of a single residential user.
n	50	Number of residential users.
$[\omega^T_{min}, \omega^T_{max}]$	[1, 3.5]	Residential users' utility parameters for traditional energy.
$[\omega^N_{min}, \omega^N_{max}]$	[1, 4]	Residential users' utility parameters for wind energy.
$[\alpha_{min}, \alpha_{max}]$	[0.5 ,1]	Coefficient of marginal utility diminishing.
$[L^T_{min}, L^T_{max}]$	[0, 200] kW	Output bounds of the thermal power unit.
$[L^N_{min}, L^N_{max}]$	[0, 200] kW	Output bounds of the wind turbine.
θ, η	0.05, 0	Cost coefficients of wind power generation.
a, b, c	0.0005, 0.25, 0	Cost coefficients of thermal power generation.
p_c	0.036	Carbon tax price per unit of carbon emission.
r, s, t	0.0034, -0.38, 36	Carbon emission coefficients of thermal power generation.
$[p_{min}, p_{max}]$	[0.1, 1.5] Yuan/kWh	Price bound for both traditional and renewable energy.

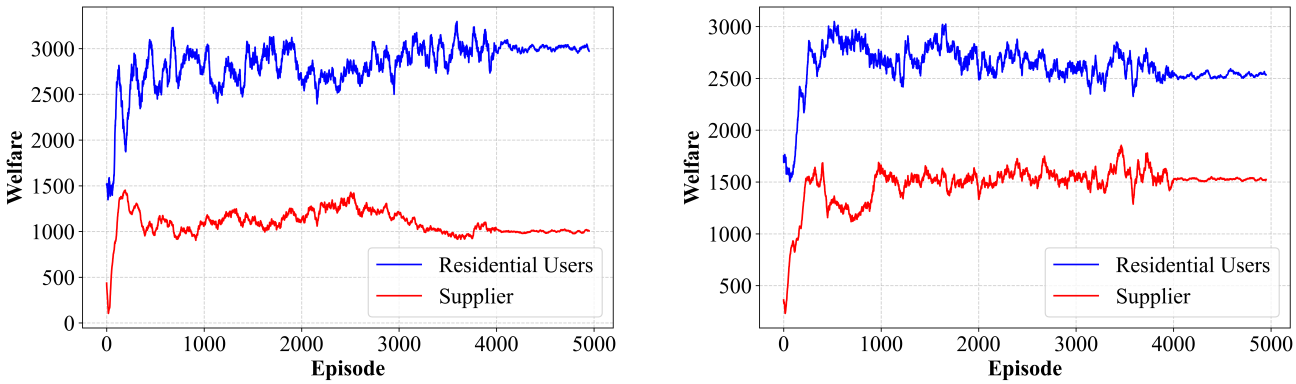
Table 2. Summary of training setup parameters.

Parameters	Value
Actor network learning rate	1×10^{-4}
Critic network learning rate	1×10^{-3}
Number of episodes	5×10^3
Discount factor	0.95
Buffer size	2×10^5
Batch size	256
Initial noise rate	0.6
Final noise rate	0.05
Actor neural network layers	4
Critic neural network layers	4
Hidden units per layer	64
Soft update coefficient	0.01

Figure 3 presents the evolution of the system reward across the training and evaluation phases of the MARL model. During the training phase, the agents progressively improve their decision-making through exploration and policy updates, leading to a rapid increase and subsequent convergence of the system reward. In the evaluation phase, the system reward remains stable, indicating that the learned policies exhibit good convergence and generalization performance.

To illustrate the performance of the proposed MARL model, we compared the welfare evolution of residential users and the supplier under two scenarios: without welfare balancing and with welfare balancing. Figure 4 shows the evolution of the welfare of residential users and the supplier during the training process under the two scenarios. In the figure, the blue curve represents the welfare of residential users, while the red curve represents the welfare of the supplier. In the scenario without welfare balancing, the welfare of both parties rises rapidly during the first 100 iterations. However,

after this initial phase, the supplier's welfare gradually declines while the welfare of residential users continues to increase. This phenomenon reflects that optimizing solely for overall social welfare often comes at the cost of one participant's welfare to maximize system-wide benefits. Around the 4000th iteration, the model reaches a stable state. In Figure 4(a), the welfare of residential users stabilizes at approximately 3039.4325, the supplier's welfare stabilizes at around 1028.5644, and the total social welfare is roughly 4067.9969. This trend indicates that neglecting welfare balancing leads to uneven welfare distribution despite achieving high overall efficiency. In Figure 4(b), the welfare of both parties also increases rapidly in the early training phase. As training progresses, the agents gradually learn better strategies, and the growth of welfare slows down, with the system converging around the 4000th iteration. At convergence, the welfare of residential users stabilizes at approximately 2529.2340, the supplier's welfare stabilizes at about 1528.3247, and the total social welfare is roughly 4057.5587. Without welfare balancing, the welfare gap between suppliers and users is approximately 2010.8681. After introducing the welfare-balancing mechanism, the gap is reduced to about 1000.9093, representing an absolute reduction of 1009.9588 and a relative decrease of approximately 50.23%. These results demonstrate that, in both scenarios, the two types of agents exhibit good convergence characteristics, indicating that the learned policies are stable and consistent within the system. Compared with the model without welfare balancing, introducing appropriately weighted welfare balancing terms in the reward function can significantly reduce the welfare gap between the supplier and the users while maintaining nearly the same total social welfare.



(a) Without welfare balancing

(b) With welfare balancing

Figure 4. Comparison of welfare convergence curves for residential users and the supplier with and without welfare balancing.

Figure 5 illustrates the evolution of the daily average prices of traditional and renewable energy during the MARL training process. In the early stages of training, both prices exhibit significant fluctuations, indicating that the agents are in an intensive exploration phase. Notably, the renewable energy price shows a relatively long period of apparent stability between approximately the 200 th and 1200 th iterations, suggesting that the agent may have temporarily fallen into a local optimum, which is closely related to the intermittency and randomness of renewable generation. As training progresses, the agents gradually learn more stable pricing strategies, and both price curves ultimately converge. The final results show that the daily average price of traditional energy stabilizes at around 1 yuan, while the renewable energy price converges to approximately 0.76 yuan.

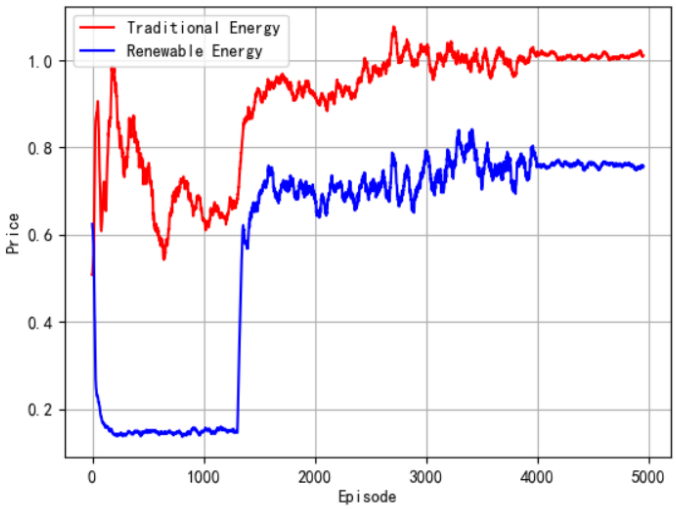


Figure 5. Convergence curves of traditional and renewable energy prices during training.

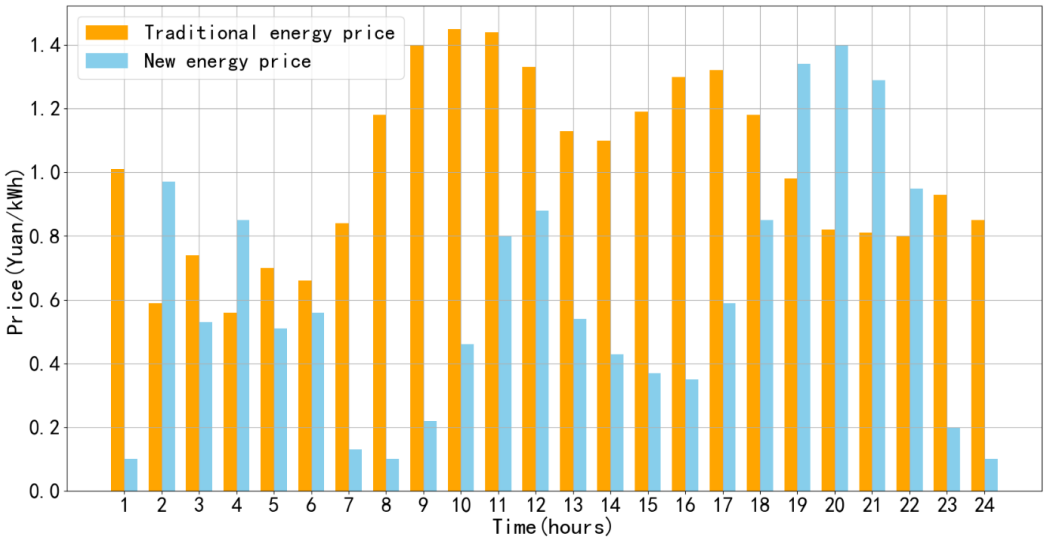
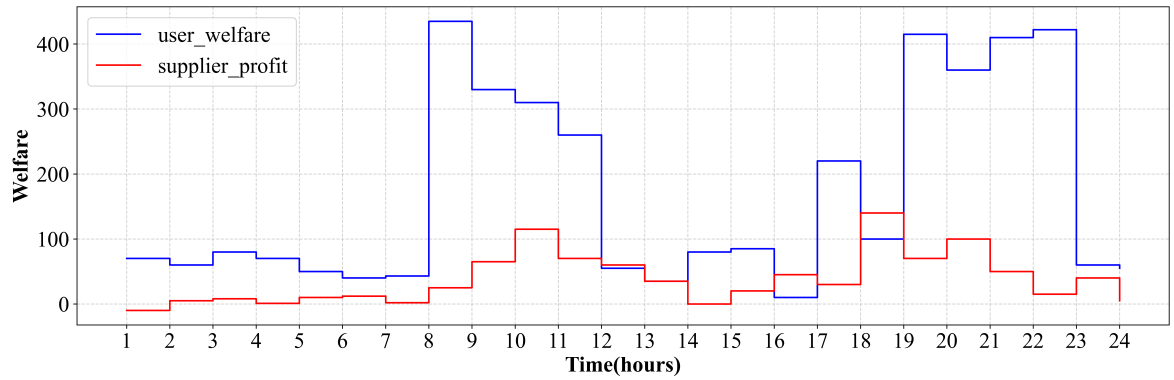


Figure 6. Hourly variation trends of traditional and new energy prices.

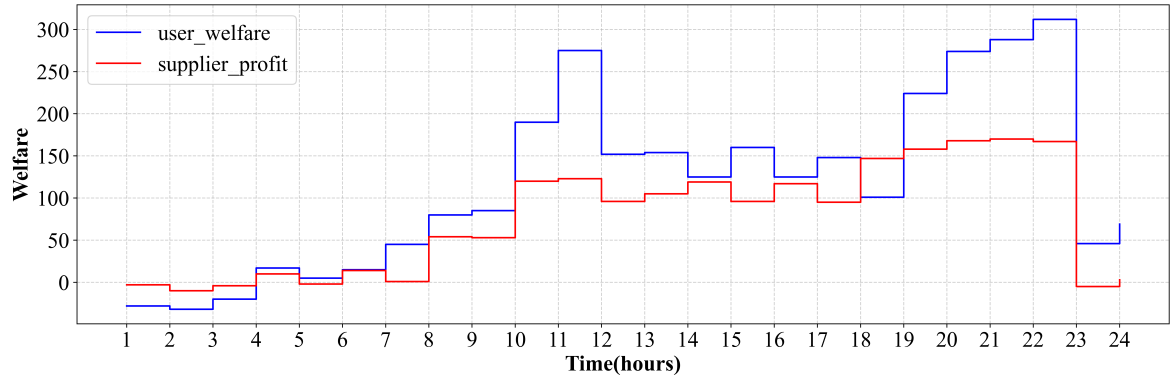
More detailed insight can be obtained by examining the stepwise time-of-use prices within a day. Figure 6 presents the stepwise time-of-use electricity prices for traditional and renewable energy after the model has converged to a stable state. The horizontal axis represents time (in hours), and the vertical axis represents electricity price (in yuan/kWh). In the figure, the orange bars indicate traditional energy prices, while the blue bars represent renewable energy prices. As shown, between hours 2 and 6, the two prices are very close, indicating low overall demand and limited renewable energy supply during this period. During daytime, the traditional energy price fluctuates in accordance with typical residential load patterns, while the renewable energy price remains relatively low, reflecting the model's incentive mechanism that encourages users to prioritize renewable energy when solar generation is abundant.

In the evening, as renewable generation decreases, its price gradually rises, prompting users to shift toward traditional energy. After 23:00, as overall demand declines, the renewable energy price also decreases accordingly.

We also analyze the welfare and profit outcomes of residential users and the supplier over a 24-hour horizon. Figure 7 shows the hourly results of both parties after model convergence, with Figure 7(a),(b) corresponding to the scenarios without and with the welfare-balancing mechanism, respectively. As shown in Figure 7(a), residential users obtain a daily welfare of 2897.7220, while the supplier achieves a profit of 1014.8798, resulting in a total social welfare of 3912.6017 and a welfare gap of 1882.8422 between the two parties. In contrast, Figure 7(b) shows that the daily welfare of residential users and the supplier are 2528.5754 and 1527.9595, respectively, and the total social welfare increases to 4056.5349, while the welfare gap is reduced to 1000.6159. Compared with the model without welfare balancing, the proposed method reduces the welfare gap between users and the supplier by approximately 46.9%. Notably, while significantly improving the fairness of welfare distribution, the total social welfare does not decrease but instead shows a slight increase, indicating that the welfare-balancing mechanism helps guide the multi-agent system toward more coordinated and efficient decision-making strategies.



(a) Without welfare balancing



(b) With welfare balancing

Figure 7. Comparison of hourly welfare curves of residential users and the supplier with and without welfare balancing.

To verify the superiority of the multi-source power supply MARL architecture proposed in this paper, we constructed a single-source power supply RL environment and adopted a deep deterministic policy

gradient (DDPG), a classic deep RL benchmark algorithm, for comparative experiments. Figure 8 presents a comparative analysis of DDPG and our proposed MADDPG model over a 24-hour simulation period in terms of user welfare and supplier profit. In the benchmark model, residential users achieve a daily welfare of 2683.3421, while the power supplier attains a daily welfare of 230.2880, with a total social welfare of 2913.6301. In comparison with the benchmark model, the proposed model in this paper increases the total social welfare by approximately 39.2%. This validates the low-cost advantage of our model in new energy integration. Furthermore, compared with heuristic optimization algorithms, the algorithm proposed in this paper has significant advantages in real-time response, privacy protection, and other key dimensions.

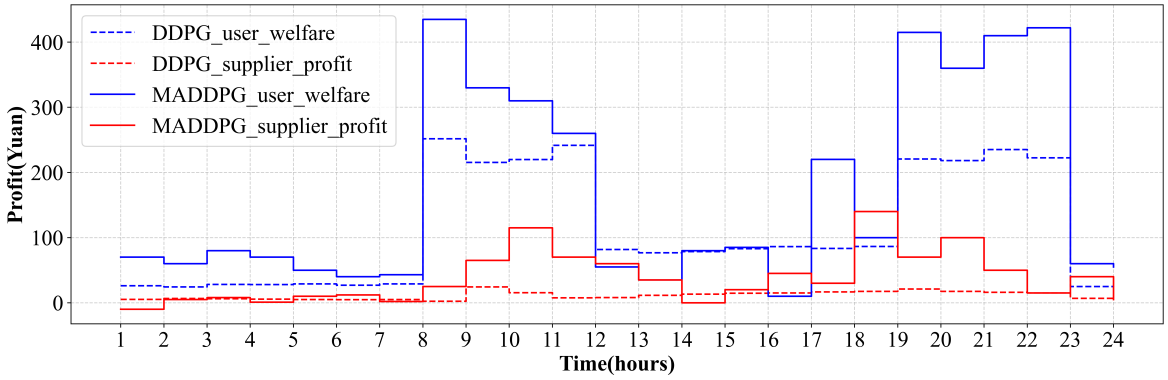


Figure 8. Hourly variation trends of user welfare and supplier profit under DDPG and MADDPG.

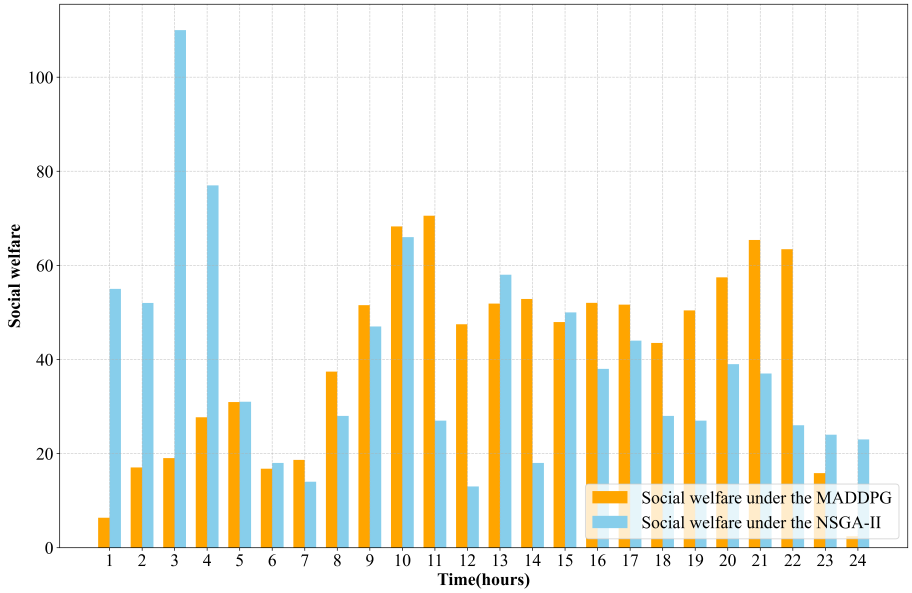


Figure 9. Hourly variation trends of social welfare under MADDPG and NSGA-II.

To further validate the effectiveness of the proposed algorithm, we compare the results obtained under the same parameter settings as those used in reference [10]. Figure 9 shows the social welfare of both approaches over a 24-hour period. The NSGA-II algorithm achieves a total welfare of 960.0440,

while the multi-agent deep deterministic policy gradient algorithm with welfare balancing achieves a total welfare of 966.5362. The small difference between the two results indicates the effectiveness of the proposed method under identical parameter settings, and the slight difference may be attributed to the inherent uncertainty of renewable energy generation.

6. Conclusions

This work develops a RTP model based on a MARL framework using the MADDPG algorithm, aiming to capture the strategic and dynamic interactions among multiple participants in the electricity market. Unlike existing research, traditional electricity market pricing models often simplify the market structure into a two-party game between suppliers and users, primarily focusing on the supplier's pricing capabilities while overlooking or underestimating supply constraints caused by the stochastic nature of renewable energy. Meanwhile, user-side responses are frequently modeled as static demand functions, making it difficult to reflect the dynamic strategic interactions between supply and demand.

Numerical simulation results demonstrate that the proposed MARL–based RTP method can effectively alleviate the excessive welfare disparity between the supply side and the demand side while improving overall social welfare. Without the introduction of the welfare balancing mechanism, the system is able to achieve a certain level of social welfare; however, a pronounced welfare imbalance persists between suppliers and residential users. After incorporating the welfare balancing mechanism, the welfare gap between the two parties is significantly reduced, while the total system welfare does not decline. These results indicate that the welfare balancing mechanism can guide multiple agents to form more coordinated pricing and electricity consumption decisions without sacrificing overall efficiency, thereby achieving a synergistic improvement in fairness and efficiency. Furthermore, comparative results with the NSGA-II multi-objective optimization algorithm show that the proposed method achieves superior or comparable performance in terms of overall welfare, while exhibiting stronger online decision-making capability and environmental adaptability. However, the present work does not incorporate energy storage facilities into the system, resulting in incomplete utilization of renewable energy during periods of low user demand. Future research could integrate electricity storage systems to achieve 100% renewable energy absorption. Furthermore, to improve the engineering practicability of the proposed pricing model, we will embed the core physical constraints, such as voltage fluctuations and line losses, into the design of the state space, action space and reward function of the MARL framework, thereby realizing the deep integration of market-driven pricing decisions and the physical operation safety of the power grid. Moreover, the current study focuses solely on residential users and overlooks other categories of electricity consumers; subsequent work may explore differentiated demand patterns across various user types. In addition, the present model considers carbon emission costs only for conventional energy. Future research could more fully account for the differences between renewable and conventional energy sources by incorporating government subsidies for renewable energy and pollution control costs.

Use of AI tools declaration

This article declares that no artificial intelligence (AI) tools were used in the writing of this article.

Acknowledgments

This work was supported by National Science Foundation of China (No.12101198) and Henan province science and technology research project (No.252102210106).

Conflict of interest

The authors of this article declare there are no conflicts of interest.

References

1. K. Dheeraja, R. Padma Priya, T. Ritika, Optimal real-time pricing and sustainable load scheduling model for smart homes using Stackelberg game theory, in *Proceedings of International Conference on Computational Intelligence and Data Engineering*, Springer, Singapore, **99** (2022). https://doi.org/10.1007/978-981-16-7182-1_22
2. Y. Dai, Y. Gao, H. Yin, X. Feng, Research on integrated demand response mechanism of electricity market considering renewable energy subsidies, *Chin. J. Manage. Sci.*, **33** (2025), 357–368. https://doi.org/10.16381/j.cnki.issn1003-207x.2022.2441
3. T. Namerikawa, N. Okubo, R. Sato, Y. Okawa, M. Ono, Real-time pricing mechanism for electricity market with built-in incentive for participation, *IEEE Trans. Smart Grid*, **6** (2015), 2714–2724. https://doi.org/10.1109/TSG.2015.2447154
4. Y. Dai, X. Sun, L. Li, H. Gao, Residential electricity real-time demand response mechanism based on multi-level game in smart grid, *Oper. Res. Manage. Sci.*, **30** (2021), 11–17. https://doi.org/10.12005/orms.2021.0307
5. P. Samadi, A. H. Mohsenian-Rad, R. Schober, V. W. S. Wong, J. Jatskevich, Optimal real-time pricing algorithm based on utility maximization for smart grid, in *2010 First IEEE International Conference on Smart Grid Communications*, Gaithersburg, MD, USA, (2010), 415–420. https://doi.org/10.1109/SMARTGRID.2010.5622077
6. G. Yuan, Y. Gao, H. Wang, A real-time pricing algorithm based on utility classification in a smart grid, *J. Univ. Shanghai Sci. Technol.*, **42** (2020), 29–35. https://doi.org/10.13255/j.cnki.jusst.2020.01.006
7. L. Song, G. Sheng, A nonsmooth Levenberg–Marquardt method based on KKT conditions for real-time pricing in smart grid, *Int. J. Electr. Power Energy Syst.*, **162** (2024), 110235. https://doi.org/10.1016/j.ijepes.2024.110235
8. H. Wang, Y. Gao, Real-time pricing method for smart grids based on complementarity problem, *J. Mod. Power Syst. Clean Energy*, **7** (2019), 1280–1293. https://doi.org/10.1007/s40565-019-0508-7
9. Y. Li, J. Li, Z. Yu, J. Dong, T. Zhou, A cosh-based smoothing Newton algorithm for the real-time pricing problem in smart grid, *Int. J. Electr. Power Energy Syst.*, **135** (2022), 107296. https://doi.org/10.1016/j.ijepes.2021.107296
10. L. Song, Y. Du, A real-time pricing dynamic algorithm for a smart grid with multi-pricing and multiple energy generation, *Electron. Res. Arch.*, **33** (2025), 2989–3006. https://doi.org/10.3934/era.2025131

11. Y. Xu, J. Han, Z. Yin, Q. Liu, C. Dai, Z. Ji, Voltage and reactive power-optimization model for active distribution networks based on second-order cone algorithm, *Computers*, **13** (2024), 95. https://doi.org/10.3390/computers13040095
12. M. A. L. Silva, S. R. de Souza, M. J. F. Souza, A. L. C. Bazzan, A reinforcement learning-based multi-agent framework applied for solving routing and scheduling problems, *Expert Syst. Appl.*, **131** (2019), 148–171. https://doi.org/10.1016/j.eswa.2019.04.056
13. A. K. Shakya, G. Pillai, S. Chakrabarty, Reinforcement learning algorithms: a brief survey, *Expert Syst. Appl.*, **231** (2023), 120495. https://doi.org/10.1016/j.eswa.2023.120495
14. R. Lu, S. H. Hong, X. Zhang, A dynamic pricing demand response algorithm for smart grid: reinforcement learning approach, *Appl. Energy*, **220** (2018), 220–230. https://doi.org/10.1016/j.apenergy.2018.03.072
15. J. Wang, Y. Gao, R. Li, Reinforcement learning based bilevel real-time pricing strategy for a smart grid with distributed energy resources, *Appl. Soft Comput.*, **155** (2024), 111474. https://doi.org/10.1016/j.asoc.2024.111474
16. H. Song, Z. Wang, Y. Gao, Bi-level real-time pricing model in multitype electricity users for welfare equilibrium: a reinforcement learning approach, *J. Renewable Sustainable Energy*, **17** (2025), 015501. https://doi.org/10.1063/5.0242836
17. M. Ahrarinouri, M. Rastegar, A. R. Seifi, Multiagent reinforcement learning for energy management in residential buildings, *IEEE Trans. Ind. Inf.*, **17** (2021), 659–666. https://doi.org/10.1109/TII.2020.2977104
18. B. C. Lai, W. Y. Chiu, Y. P. Tsai, Multiagent reinforcement learning for community energy management to mitigate peak rebounds under renewable energy uncertainty, *IEEE Trans. Emerging Top. Comput. Intell.*, **6** (2022), 568–579. https://doi.org/10.1109/TETCI.2022.3157026
19. Y. He, C. Gu, Y. Gao, J. Wang, Bi-level day-ahead and real-time hybrid pricing model and its reinforcement learning method, *Energy*, **322** (2025), 135316. https://doi.org/10.1016/j.energy.2025.135316
20. Y. Du, F. Li, H. Zandi, Y. Xue, Approximating Nash equilibrium in day-ahead electricity market bidding with multi-agent deep reinforcement learning, *J. Mod. Power Syst. Clean Energy*, **9** (2021), 534–544. https://doi.org/10.35833/MPCE.2020.000502
21. T. A. Nakabi, P. Toivanen, Deep reinforcement learning for energy management in a microgrid with flexible demand, *Sustainable Energy Grids Networks*, **25** (2021), 100413. https://doi.org/10.1016/j.segan.2020.100413
22. Y. Wang, H. Zhang, Y. An, Z. Ji, I. Ganchev, RG hyperparameter optimization approach for improved indirect prediction of blood glucose levels by boosting ensemble learning, *Electronics*, **10** (2021), 1797. https://doi.org/10.3390/electronics10151797
23. Y. Chen, J. Xu, Solar and wind power data from the Chinese State Grid Renewable Energy Generation Forecasting Competition, *Sci. Data*, **9** (2022), 577. https://doi.org/10.1038/s41597-022-01696-6

24. Q. Zhang, Y. Sun, Real-time pricing model of smart grid under dual-carbon target based on social welfare maximisation, *Econ. Comput. Econ. Cybern. Stud. Res.*, **58** (2024), 299–313. https://doi.org/10.24818/18423264/58.2.24.18

25. N. Harder, R. Qussous, A. Weidlich, Fit for purpose: modeling wholesale electricity markets realistically with multi-agent deep reinforcement learning, *Energy AI*, **14** (2023), 100295. https://doi.org/10.1016/j.egyai.2023.100295

26. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., Attention is all you need, in *Advances in Neural Information Processing Systems*, **30** (2017).

27. R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, I. Mordatch, Multi-agent actor-critic for mixed cooperative-competitive environments, preprint, arXiv:1706.02275. https://doi.org/10.48550/arXiv.1706.02275

AIMS Press