Synergistic Carbon Trading and Power Generation Decision Considering the Annual Compliance Cycle and Market Response: a Hybrid Mathematical-deep Reinforcement Learning Optimization Approach

Shouyuan Shi, Zhenning Pan, *Member*, *IEEE*, Junbin Chen, and Tao Yu, *Senior Member*, *IEEE*

***Abstract*—The annual compliance cycle of the carbon trading system allows generation companies (GenCos) to decouple the timing of carbon allowance purchases from their actual emissions. However, trading a large volume of allowances within a single day can significantly impact on carbon prices. Faced with uncertain future carbon and electricity prices, GenCos must address a challenging multistage stochastic optimization problem to coordinate their carbon trading strategies with daily power generation decisions. In this paper, a two-layered hybrid mathematical-deep reinforcement learning (DRL) optimization framework is proposed. The upper DRL layer tackles the stochastic, year-long carbon trading and allowance usage optimization problem, aiming for long-term optimality and providing guidance for short-term decisions in the lower layer. The lower mathematical optimization layer addresses the deterministic daily power generation schedule problem while enforcing strict technical constraints. To accelerate learning of the annual compliance cycle, a decision timeline transfer learning method is proposed, enabling the DRL agent to progressively refine its policy through sequentially training on monthly, weekly and daily decision environments. Case studies demonstrate that, with these methods, a GenCo can reduce emission costs and increase profits by effectively leveraging carbon price fluctuations within the compliance cycle.**

***Index Terms*—Carbon trading market, deep reinforcement learning, electricity market, generation company, market response.**

Received: August 6, 2024
Accepted: September 26, 2024
Published Online: January 1, 2026
Shouyuan Shi, Zhenning Pan, Junbin Chen, and Tao Yu (corresponding author) are with School of Electric Power Engineering, South China University of Technology, Guangzhou 510641, China (e-mail: shishouyuan@outlook.com; scutpanzn@163.com; junbin0617@outlook.com; taoyu1@scut.edu.cn).
DOI:10.23919/PCMP.2024.000327

NOMENCLATURE

A. Abbreviations

CTS	carbon trading system
DRL	deep reinforcement learning
DTTL	decision timeline transfer learning
GHG	greenhouse gas
GenCo	generation company
HMDRL	hybrid mathematical-deep reinforcement learning decision framework
RL	reinforcement learning
SCTPGD	synergistic carbon trading and power generation decision
SPTD	speculative and productive trade decomposition

I. INTRODUCTION

As a market-based instrument, carbon trading systems (CTS) have been implemented in many economies, most prominently in the EU and China, to mitigate greenhouse gas (GHG) emissions. Unlike carbon taxing which has a fixed fee rate, the carbon price in a CTS is driven by the market and varies over time [1], [2]. As major emitters, generation companies (GenCos) need to make synergistic carbon trading and power generation decisions (SCTPGD), which are subject to uncertainties in both carbon and electricity markets.

Unlike electricity that is largely produced and consumed simultaneously, carbon emissions and allowances do not require instantaneous balancing. As shown in Fig. 1, CTS operates on an annual compliance cycle, giving GenCos the flexibility to time their allowance purchases throughout the year, as long as they can surrender a sufficient number of allowances by the deadline in the following year [1], [2]. While fluctuating carbon prices create opportunities for GenCos to purchase allowances at lower prices with an effective

strategy, the annual compliance cycle also adds complexity to their decision-making. First, with uncertain future prices, GenCos must determine the best allowance trading time and quantity in a year. Second, when emissions and allowance are not purchased on the same day, GenCos must determine the quantity of allowances to allocate for that day's generation (allowance usage). Finally, the annual compliance cycle allows GenCos to purchase large quantities of allowances in a single day, which can increase the demand in the carbon market and raise the carbon price, as shown in Fig. 1. Therefore, it is important to take into account the carbon market's responses to GenCo's actions when the annual compliance cycle is considered.

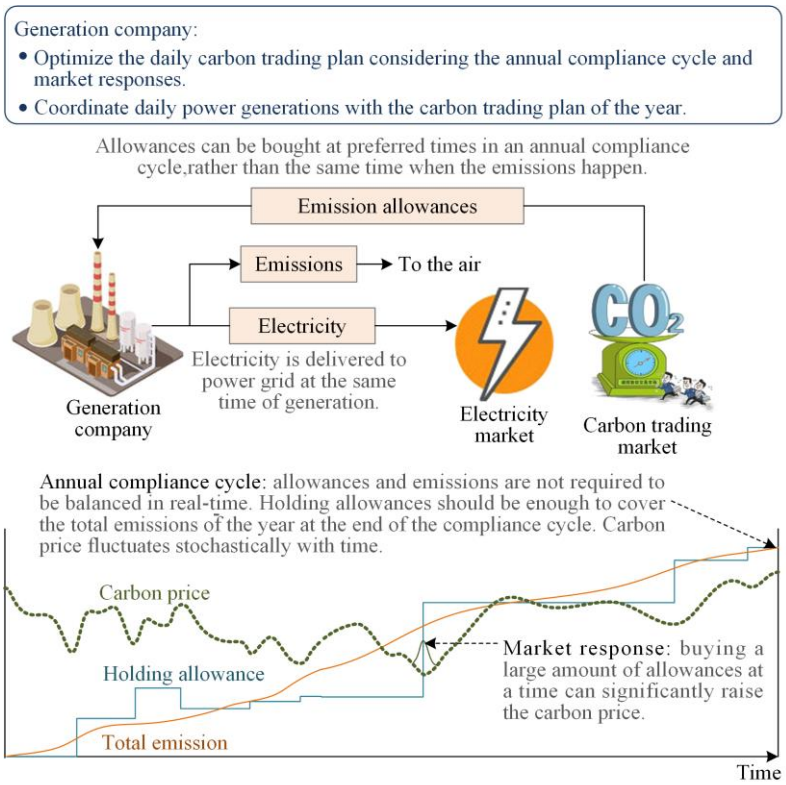


Fig. 1. Daily synergistic carbon trading and power generation decision of a GenCo considering the annual compliance cycle and market response. The end of a compliance cycle is assumed to be the last day of the year.

To the best of our knowledge, no existing study has jointly considered the annual compliance cycle and carbon market responses within the GenCo's decision-making problem. Most existing studies, such as [3]–[8], ignore both the annual compliance cycle and the market responses of the CTS. They typically assume that GenCos purchase allowances at a known price on the same day the emissions occur, thereby limiting the flexibility of the trading strategies. The carbon market response has been considered in some studies [9]–[13]. Although piece-wise linear functions are employed in [9]–[12] to model the positive correlation between carbon prices and allowance purchase volumes, while [13] further incorporates daily bidding and clearing processes of both electricity and carbon markets, none of these studies considers the annual compliance cycle. Some studies [14]–[17] incorporate the annual compliance cycle, though [14] only considers uncertainties in the electricity market and treats the carbon price as a deterministic value. Both future carbon and electricity price uncertainties are considered in a two-stage stochastic optimization model in [15], but the study only addresses monthly decision-making without incorporating daily decisions. The concept of hierarchical decision-making is proposed in [16], where a GenCo sequentially makes decisions at yearly, monthly, weekly, and daily levels to coordinate long-term and short-term interests within the annual compliance cycle. However, the study focuses primarily on short-term problems, with the case study analyzing only the final week of the compliance cycle. The optimal stopping theory is used in [17] to choose the optimal allowance buying time in daily level under carbon price uncertainty, but power generation decisions are not considered, and the emission amount is assumed to be known without uncertainty. None of these studies considering the annual compliance cycle has taken the carbon market responses into account. Although our previous work [18] employs dynamic programming to solve the 365 daily carbon trading and generation problems within the annual

compliance cycle, the GenCo is assumed to be a price-taker while carbon market responses are omitted.

Addressing the daily SCTPGD problem with the annual compliance cycle and market responses is inherently challenging. First, due to the annual compliance cycle, GenCos must account for the entire year in their planning, as daily decisions are temporally interdependent. In fact, the SCTPGD problem is a multistage stochastic optimization problem with 365 stages (366 for leap years). There will be totally 2^{365-1} paths in the scenario tree even only 2 possible situations are considered for each day. Thus, it is intractable for the traditional scenario tree based stochastic optimization method [19]. If carbon market responses are not considered, a GenCo can make its carbon trading decisions by solely optimizing the trading time for a given number of emissions. This simplified setting is adopted by [17] and our previous work in [18], and consequently, the methods are not applicable when carbon market responses need to be incorporated.

Deep reinforcement learning (DRL) is a powerful methodology for solving multistage stochastic optimization problems. By integrating reinforcement learning (RL) with deep learning (DL), DRL can improves its decision strategy by interacting with the environment through the RL mechanism, and learn complicated behaviors with the powerful function approximation capability of neural networks. Existing studies have applied DRL in power generation and carbon trading decisions [8], [9], [20]. However, these studies focus exclusively on single-day decision-making and overlook the annual compliance cycle, which diverges from real-world practices. Consequently, the application of DRL to the SCTPGD problem within an annual compliance cycle remains unexplored. A key limitation of DRL is its difficulty in handling strict and complex constraints, such as generator operational limits, precisely the areas where mathematical optimization methods excel. Another challenge is that the SCTPGD problem presents a sparse reward environment for DRL [21], since the agent receives the allowance shortage penalty only once a year (every 365 steps), resulting in slow learning convergence.

In this paper, the daily SCTPGD problem of a GenCo is addressed by integrating DRL with mathematical optimization method to maximize the GenCo's total annual profit. The main contributions include:

1) Problem formulation: The daily SCTPGD problem of a GenCo is formulated as a multistage stochastic optimization problem that jointly incorporates the annual compliance cycle and carbon market responses. This formulation enables the GenCo to reduce emission costs by strategically timing allowance purchases throughout the year, while avoiding large single-day transactions that might drive up carbon prices. Additionally, carbon market speculation activities are accounted for in the model;

2) Decision framework: A two-layered hybrid mathematical-deep reinforcement learning (HMDRL) decision framework is proposed to solve the SCTPGD problem. The upper DRL layer handles the multistage stochastic optimization of carbon trading and allowance usage problem, pursuing long-term optimality and providing daily allowance usage suggestion for short-term decisions in the lower layer. The lower mathematical optimization layer handles the deterministic daily power generation optimization problem with strict technical constraints, maximizing daily electricity selling profits. With the HMDRL framework, the technical constraints of the generators are strictly followed, and the lower layer can be easily extended to include more decision elements of the GenCo, since most of the studies in this area are based on mathematical optimization methods;

3) DRL techniques: A decision timeline transfer learning (DTTL) method is proposed to mitigate the sparsity of allowance shortage penalty. The DRL agent is first trained sequentially on monthly and weekly decision timelines, enabling it to quickly learn the annual compliance cycle with fewer steps. It is then trained on a daily decision timeline to refine its policy. Additionally, modifications to the DRL action space and the critic network's input structure are proposed to address learning failures that occur when allowance selling is prohibited;

4) Behavior analyses: First, numerical studies are conducted to examine how the annual compliance cycle and market responses influence the GenCo's operational decisions. Second, when the GenCo is permitted to engage in carbon market speculation, its allowance holdings may fluctuate multiple times throughout the year. To better analyze the GenCo's behavior, an ex-post speculative and productive trade decomposition (SPTD) algorithm is proposed to distinguish between allowances acquired for speculative purposes and those for power generation.

II. Problem Formulation

A. Scope and Assumptions

As shown in Fig. 1, this paper addresses the daily SCTPGD problem of a GenCo, taking into account the annual compliance cycle and market response of the CTS. For each decision day, the SCTPGD problem involves determining the volume of allowances to be traded and the amount of electricity to be generated per hour.

To focus on the synergy between daily generation decisions with the carbon trading plan of the whole year, only inter-day uncertainties are considered. Intra-day prices are assumed to be known, and real-time operational

adjustments within a day are not considered. The GenCo studied in this paper is assumed to be a price-taker in the electricity market, and the electricity market bidding and clearing processes are ignored. The bidding curve can be constructed from the desired generation curve using the method in [22]. The participation in the ancillary service market and long-term bilateral electricity contracts is also out of the scope of this paper.

It is assumed that allowances are surrendered on the last day of the current year, and any surplus allowances are disregarded, as inter-year coordination is beyond the scope of this paper. The influence of the GenCo's carbon trading actions on carbon prices is considered, since purchasing a full year's allowances within just a few days could lead to substantial single-day transactions.

It is infeasible to train the DRL agent directly in the real world, as evaluating its performance over a single episode requires a full year of interactions due to the annual compliance cycle. Therefore, the agent must be trained in a virtual environment that simulates market prices and responses, which can be constructed by market analysts.

B. Formulation of the SCTPGD Problem

To leverage the annual compliance cycle, the GenCo must optimize both current-day carbon-electricity operations and anticipate future-day decisions, ensuring daily actions align with long-term strategy. Although it solves a multistage stochastic optimization problem each day, only the current day's decisions are implemented.

1) Objective

The objective of the SCTPGD problem is to maximize the overall profit within an annual compliance cycle by optimizing carbon trading amount and power outputs of the generators every day as:

$$\max_{\substack{\boldsymbol{p}_d, v_d \\ d=d_0}} \left[\begin{array}{l} r_d(\boldsymbol{p}_d, v_d, \boldsymbol{\lambda}_{\mathrm{E},d}, \lambda_{\mathrm{C}_0,d}) + \\ \mathop{\mathbb{E}}_{\substack{\boldsymbol{\lambda}_{\mathrm{E},d+1} \\ \lambda_{\mathrm{C}_0,d+1}}} \left[\begin{array}{l} \max\limits_{\boldsymbol{p}_{d+1}, v_{d+1}} r_{d+1}(\boldsymbol{p}_{d+1}, v_{d+1}, \boldsymbol{\lambda}_{\mathrm{E},d+1}, \lambda_{\mathrm{C}_0,d+1}) + \cdots + \\ \mathop{\mathbb{E}}\limits_{\substack{\boldsymbol{\lambda}_{\mathrm{E},D} \\ \lambda_{\mathrm{C}_0,D}}} \left[\max\limits_{\boldsymbol{p}_D, v_D} r_D(\boldsymbol{p}_D, v_D, \boldsymbol{\lambda}_{\mathrm{E},D}, \lambda_{\mathrm{C}_0,D}) \right] \cdots \end{array} \right] \end{array} \right] \tag{1}$$

$$r_{d|d<D}(\cdot) = \left[\begin{array}{l} \sum\limits_h \sum\limits_g \left[\begin{array}{l} \left(\lambda_{\mathrm{E},d,h} p_{d,h,g} - \lambda_{\mathrm{W},g} w_{d,h,g} \right) \Delta h - \\ \lambda_{\mathrm{U},g} u_{\mathrm{U},d,h,g} - \lambda_{\mathrm{D},g} u_{\mathrm{D},d,h,g} \end{array} \right] - \\ \lambda_{\mathrm{C},d} v_d \end{array} \right] \tag{2}$$

$$r_D(\cdot) = \left[\begin{array}{l} \sum\limits_h \sum\limits_g \left[\begin{array}{l} \left(\lambda_{\mathrm{E},d,h} p_{D,h,g} - \lambda_{\mathrm{W},g} w_{D,h,g} \right) \Delta h - \\ \lambda_{\mathrm{U},g} u_{\mathrm{U},D,h,g} - \lambda_{\mathrm{D},g} u_{\mathrm{D},D,h,g} \end{array} \right] - \\ \lambda_{\mathrm{C},D} v_D - \lambda_{\mathrm{pen}} [E_D - V_D]^+ \end{array} \right] \tag{3}$$

where d, h and g are the day index, hour index and generator index respectively; d_0 is the index of the current day; $d \in [1, D]$; and D is the index of the last day of the year, which is 365 for a normal year or 366 for a leap year; $\boldsymbol{p}_d$ is the vector of $p_{d,h,g}$, i.e., the output power of generator g in day d hour h; v_d is the amount of allowances traded in day d (positive for buying and negative for selling); $\boldsymbol{\lambda}_{\mathrm{E},d}$ is the vector of hourly electricity price $\lambda_{\mathrm{E},d,h}$ in day d; $\lambda_{\mathrm{C}_0,d}$ is the original carbon price in day d before affected by the GenCo's action; and $\lambda_{\mathrm{C},d}$ is the actual carbon price in day d affected by the GenCo's action shown in (48); $\lambda_{\mathrm{W},g}$ is the fuel price of generator g; and $w_{d,h,g}$ is the fuel consumption rate that are specified in (11); Δh is the duration of an hour; while $\lambda_{\mathrm{U},g}$ and $\lambda_{\mathrm{D},g}$ are the startup and shutdown cost of generator g, respectively; $u_{\mathrm{U},d,h,g}$ and $u_{\mathrm{D},d,h,g}$ are binary variables that indicate whether or not the generator is started or stopped in the corresponding time step; while E_D and V_D are respectively the yearly total emissions and the allowances held by the GenCo at the end of the compliance cycle; $[\cdot]^+ = \max\{0, \cdot\}$; and $[E_D - V_D]^+$ is the allowance shortage of the year; λ_{pen} is the penalty price for a unit of allowance shortage; and r_d is the daily profit in day d. For $d < D$, the daily profit consists of the electricity selling revenue, the generation cost and the allowances buying cost as in (2); For $d = D$, the penalty cost for allowance shortage is added in (3) in addition to the terms in (2).

2) Operational Constraints of Generators

Generators must follow the operational constraints:

$$u_{d,h,g} P_{\min,g} \leqslant p_{d,h,g} \leqslant u_{d,h,g} P_{\max,g} \tag{4}$$

$$-P_{\mathrm{D},g} \leqslant p_{d,h,g} - p_{d,h-1,g} \leqslant P_{\mathrm{U},g} \tag{5}$$

$$u_{\mathrm{U},d,h,g} + \sum\nolimits_{h'=1}^{h_{\mathrm{D},g}} u_{\mathrm{D},d,h-h',g} \leqslant 1 \tag{6}$$

$$u_{\mathrm{D},d,h,g} + \sum\nolimits_{h'=1}^{h_{\mathrm{U},g}} u_{\mathrm{U},d,h-h',g} \leqslant 1 \tag{7}$$

$$u_{d,h,g}, u_{\mathrm{U},d,h,g}, u_{\mathrm{D},d,h,g} \in \{0,1\} \tag{8}$$

$$u_{\mathrm{U},d,h,g} \geqslant u_{d,h,g} - u_{d,h-1,g} \tag{9}$$

$$u_{\mathrm{D},d,h,g} \geqslant u_{d,h-1,g} - u_{d,h,g} \tag{10}$$

$$w_{d,h,g} = \alpha_{g,1} p_{d,h,g} + \alpha_{g,0} u_{d,h,g} \tag{11}$$

$$e_d = \sum_h \sum_g \left(k_g w_{d,h,g} - k_{\mathrm{free},g} p_{d,h,g} \right) \Delta h \tag{12}$$

where $u_{d,h,g}$ is a binary variable that indicates the on/off state of the generator; $P_{\max,g}$ and $P_{\min,g}$ are the maximum and minimum power limit of generator g, respectively; while $P_{\mathrm{D},g}$ and $P_{\mathrm{U},g}$ are the per-hour maximum ramp-down and ramp-down power, respectively; $h_{\mathrm{D},g}$ and $h_{\mathrm{U},g}$ are the minimum off/on time in hours,

respectively; while $\alpha_{g,1}$ and $\alpha_{g,0}$ are the coefficients of the fuel consumption rate of generator g; e_d is the daily net emission amount; k_g is the emission coefficient; and $k_{\text{free},g}$ is the coefficient for free allowances. Equation (4) ensures that the output power of the generator falls within the allowable range, while (5) is power ramping constraint. Equations (6) and (7) ensure that the generator can only be started/stopped after a minimum time has passed since its last shutdown/startup respectively. Equation (8) requires that variables relating to the on/off state of the generator to be binary, while (9) and (10) ensure the logical relations between them. Equation (11) calculates the fuel consumption rate with the output power and on/off state of the generator, while (12) calculates the daily net emissions.

3) Carbon Trading Constraints

The GenCo needs to track the number of its emissions and holding allowances every day. Equations (13) and (14) update the number of accumulated emissions E_d and holding allowances V_d every day respectively, while (15) restricts the daily allowance trading amount to be within a given range, where v_{buy} and v_{sell} are the maximum daily buying and selling amount respectively, either set by the government or the GenCo itself.

$$E_d = E_{d-1} + e_d \tag{13}$$

$$V_d = V_{d-1} + v_d \geqslant 0 \tag{14}$$

$$-v_{\text{sell}} \leqslant v_d \leqslant v_{\text{buy}} \tag{15}$$

4) Electricity and Carbon Price Models

The study of the electricity and carbon price models is sufficiently important for it to be a dedicated research topic and thus is beyond the scope of this paper. The HMDRL decision framework does not rely on specific price models, as long as the prices can be observed from the environment, which can be described abstractly as:

$$\lambda_{\text{E},d,h} = f_{\text{E}}(\boldsymbol{\xi}_d) \tag{16}$$

$$\lambda_{\text{C},d} = f_{\text{C}}(\boldsymbol{\xi}_d, v_d) \tag{17}$$

where $\boldsymbol{\xi}_d$ is the vector of environment internal stochastic parameters that affect the electricity and carbon prices, such as the electricity demand, renewable energy outputs, the actions of other GenCos, etc. Function $f_{\text{E}}(\boldsymbol{\xi}_d)$ means that the electricity prices are determined by the stochastic parameters. Function $f_{\text{C}}(\boldsymbol{\xi}_d, v_d)$ means the carbon price is determined by both the stochastic parameters and the allowance trading amount of the GenCo, which depicts the carbon market's responses to the GenCo's actions. Functions $f_{\text{E}}(\boldsymbol{\xi}_d)$ and $f_{\text{C}}(\boldsymbol{\xi}_d, v_d)$ can be in any form, such as analytical functions, differential equations, neural networks, or even complicated simulators.

III. Hybrid Mathematical-Deep Reinforcement Learning Decision Framework

To combine the power of DRL and mathematical optimization methods, a hybrid mathematical-deep reinforcement learning decision framework is proposed as shown in Fig. 2. The SCTPGD problem is decomposed into a two-layer framework. The upper DRL layer handles the multistage stochastic carbon trading and allowance utilization to pursue long-term optimality, offering allowance usage recommendation for short-term decisions in the lower layer. The lower mathematical optimization layer addresses the single-day deterministic power generation optimization problem subject to complex operational constraints.

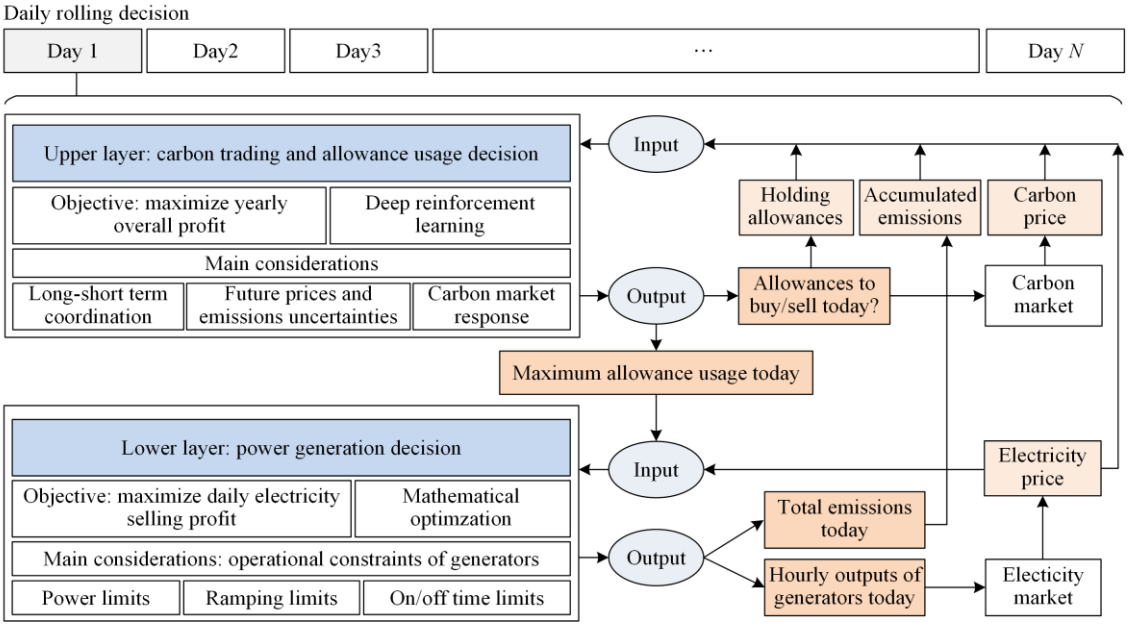


Fig. 2. The two-layer hybrid mathematical-deep reinforcement learning decision framework.

DRL is used in the upper layer since the 365-stage stochastic optimization problem is difficult to be handled by the scenario-based sample average method. Mathematical optimization is used in the lower layer for both reliability and extensibility.

1) Reliability: There are some strict technical constraints on the physical system that the GenCo must follow, which can be explicitly guaranteed by the mathematical optimization model.

2) Extensibility: Beyond the generation decision problem addressed in the lower layer of this study, many other decision problems discussed in the literature, such as strategic bidding and ancillary service market participation, are predominantly studied using mathematical optimization methods rather than DRL. These existing methods can be easily incorporated into the lower layer without modifying the upper layer structure, since the upper layer only deals with carbon trading and allowance usage decisions. In contrast, if the lower layer were implemented using DRL, the GenCo would no longer be able to leverage existing mathematical optimization based approaches, and both the structure and training techniques might need to be redesigned for the new decision problems.

A. Two-layer Decision Formulation

1) The Upper Layer

On each decision day, the upper layer DRL agent determines both the number of allowances to trade and the amount to allocate for use by the lower layer that day, with the overall goal of maximizing the GenCo's annual profit. This requires the agent to balance the current day's profit against the total profit over the annual compliance cycle. Meanwhile, the DRL agent must account for the uncertainties of future prices and emissions, as well as potential market price responses to its own carbon trading actions. The upper layer optimization problem is formulated as (18), and will be reformulated into a Markov decision process (MDP) later.

$$\max_{v_d, v_{\mathrm{A},d}} \left[\begin{array}{l} r_{\mathrm{C,d}}(v_d) + r_{\mathrm{E},d}(v_{\mathrm{A},d}) + \\ \mathbb{E}\left[\begin{array}{l} \max\limits_{v_{d+1}, v_{\mathrm{A},d+1}} r_{\mathrm{C},d+1}(v_{d+1}) + r_{\mathrm{E},d+1}(v_{\mathrm{A},d+1}) + \cdots + \\ \mathbb{E} \max\limits_{v_D, v_{\mathrm{A},D}} r_{\mathrm{C},D}(v_D) + r_{\mathrm{E},D}(v_{\mathrm{A},D}) \end{array} \right] \end{array} \right] \tag{18}$$

s.t. (13)–(15) and

$$r_{\mathrm{C},d|d<D}(v_d) = -\lambda_{\mathrm{C},d} v_d \tag{19}$$

$$r_{\mathrm{C},D}(v_D) = -\lambda_{\mathrm{C},D} v_D - \lambda_{\mathrm{pen}} [E_D - V_D]^+ \tag{20}$$

$$0 \leqslant v_{\mathrm{A},d} \leqslant v_{\mathrm{A,max}} \tag{21}$$

where $r_{\mathrm{C},d}(v_d)$ is the profit from the carbon market by trading v_d of allowances; $v_{\mathrm{A},d}$ is the amount of maximum allowance usage allocated for that day; $r_{\mathrm{E},d}(v_{\mathrm{A},d})$ is the profit from the electricity market obtained by the lower layer; and $v_{\mathrm{A,max}}$ is the maximum possible daily allowance usage, which equals to the emission amount when all generators run in full power throughout the day.

2) The Lower Layer

For each decision day, the lower layer optimize the hourly output power of each generator to maximize the daily electricity selling profit as:

$$r_{\mathrm{E},d}(v_{\mathrm{A},d}) = \max_{p_{d,h,g}} \left[\begin{array}{l} \sum\limits_h \sum\limits_g \left[\begin{array}{l} \left(\lambda_{\mathrm{E},d,h} p_{d,h,g} - \lambda_{\mathrm{W},g} w_{d,h,g}\right) \Delta h - \\ \lambda_{\mathrm{U},g} u_{\mathrm{U},d,h,g} - \lambda_{\mathrm{D},g} u_{\mathrm{D},d,h,g} \end{array} \right] - \\ \lambda_{\mathrm{pen}} [e_d - v_{\mathrm{A},d}]^+ \end{array} \right] \tag{22}$$

s.t. (4)–(12).

The lower layer does not involve decisions in the future, but will receive a penalty for emissions that exceed the daily maximum allowance usage allocated by the upper layer, to ensure that the long-term plan of the upper layer is affected. The penalty price is set as the penalty price λ_{pen} of the CTS, which is based on the concept that if the lower layer does not follow the strategy of the upper layer, it may lead to an allowance shortage in the compliance cycle.

All operational constraints of the generators are taken into account in the lower layer, so (22) is a mixed integer linear programming problem that can be solved with many off-the-shelf solvers. Since the maximum allowance usage $v_{\mathrm{A},d}$ given by the upper layer are enforced in the lower layer by penalty instead of hard constraints, the action of the upper layer will not affect the feasibility of the lower layer problem.

B. The DRL Algorithm

DRL is used in the upper layer to make daily carbon trading and allowance usage decisions from the perspective of the entire annual compliance cycle under uncertainty. Given the abundance of powerful DRL algorithms already available, this paper does not aim to propose a new one. Instead, it adopts the twin delayed deep deterministic policy gradient (TD3) algorithm, augmented with several enhancements to improve its perform on the SCTPGD problem.

Figure 3 shows the structure of the SCTPGD environment and the TD3 agent. The MDP formulation of the SCTPGD problem and the basic TD3 algorithm are introduced first in the following contents, while further enhancements to the TD3 algorithm are introduced later.

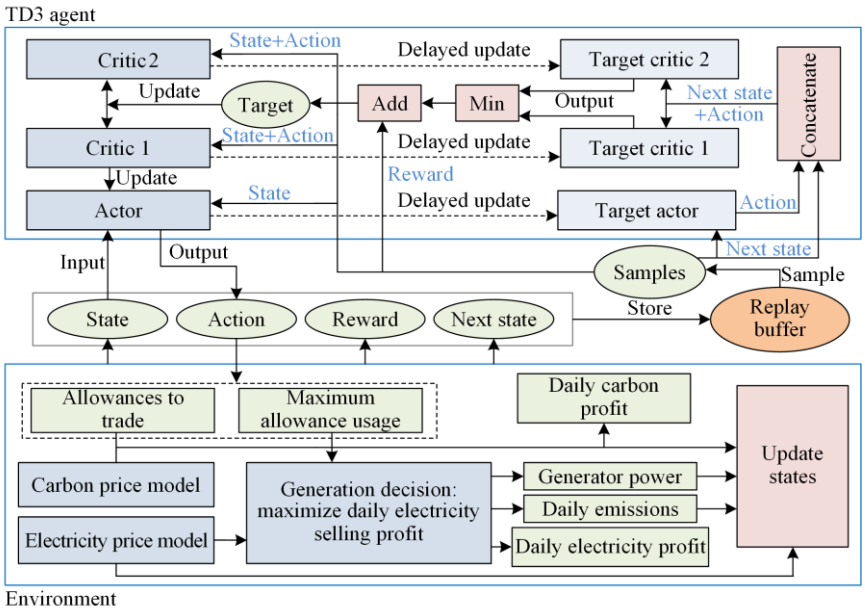


Fig. 3. TD3 algorithm and the environment.

1) MDP Formulation of the SCTPGD Problem

State space: $\boldsymbol{S}_d = (d, \lambda_{\mathrm{C}_0,d}, \overline{\lambda}_{\mathrm{E},d}, \sigma_{\mathrm{E},d}, E_{d-1}, V_{d-1})$. In order to make the upper-layer decisions, the agent needs to observe the current time d, the current carbon price $\lambda_{\mathrm{C}_0,d}$, the average value $(\overline{\lambda}_{\mathrm{E},d})$ and standard deviation $(\sigma_{\mathrm{E},d})$ of the hourly electricity prices in the day, as well as the latest values of the accumulated emissions E_{d-1} and holding allowances V_{d-1}.

Action space: $\boldsymbol{A}_d = (v_d, v_{\mathrm{A},d})$. As mentioned above, the agent determines the carbon trading amount v_d and maximum allowance usage $v_{\mathrm{A},d}$ in day d. $v_d \in \left[-v_{\mathrm{sell}}, v_{\mathrm{buy}}\right]$, $v_{\mathrm{A},d} \in \left[0, v_{\mathrm{A,max}}\right]$.

State transition: $\Pr(\boldsymbol{S}_{d+1} \mid \boldsymbol{S}_d, \boldsymbol{A}_d)$. When action $\boldsymbol{A}_d$ is applied at state $\boldsymbol{S}_d$, the environment turns to state $\boldsymbol{S}_{d+1} = (d+1, \lambda_{\mathrm{C}_0,d+1}, \overline{\lambda}_{\mathrm{E},d+1}, \sigma_{\mathrm{E},d+1}, E_d, V_d)$. The update of E_d is described in (13), while the update of V_d is as follows:

$$v'_d = \begin{cases} \max\{-V_{d-1}, v_d\}, & \text{if } d < D \\ \mathrm{clip}(E_{d-1} - V_{d-1}, -v_{\mathrm{sell}}, v_{\mathrm{buy}}), & \text{if } d = D \end{cases} \tag{23}$$

$$V_d = V_{d-1} + v'_d \tag{24}$$

where v'_d is the corrected carbon trading amount. The first line of (23) means that the GenCo cannot sell allowances more that it has. The second line of (23) indicates that, by the end of the annual compliance, the GenCo should aim to align its remaining allowances as close as possible with its accumulated emissions. $\mathrm{clip}(x, y, z)$ equals to $\max\{y, \min\{z, x\}\}$, which is a popular notation used in the DRL community.

Reward: $R_d(\boldsymbol{S}_d, \boldsymbol{A}_d) = r_{\mathrm{C},d} + r_{\mathrm{E},d} - \lambda_{\mathrm{pen}} \left| v_d - v'_d \right|$. $r_{\mathrm{C},d}$ is the carbon trading profit calculated in (19) and (20), while $r_{\mathrm{E},d}$ is the electricity selling profit calculated by the lower layer in (22). The third term $\lambda_{\mathrm{pen}} \left| v_d - v'_d \right|$ is the penalty for the difference between the agent's action and the corrected carbon trading amount in (23).

Policy: $\pi(\cdot): \boldsymbol{S}_d \to \boldsymbol{A}_d$. The policy π gives decisions $\boldsymbol{A}_d$ according to the current state $\boldsymbol{S}_d$, which is the main job of the DRL agent.

Action value function: $Q_\pi(\boldsymbol{S}_d, \boldsymbol{A}_d)$. The action value function is often named as "Q-function" in the DRL community, which calculates the expected accumulated rewards if action $\boldsymbol{A}_d$ is taken at state $\boldsymbol{S}_d$ and the policy π is followed onwards.

$$Q_\pi(\mathbf{S}_d, \mathbf{A}_d) = \mathbb{E}\left(R_d(\mathbf{S}_d, \mathbf{A}_d) + \sum_{d'=d+1}^{D} \gamma^{d'-d} R_{d'}(\mathbf{S}_{d'}, \pi(\mathbf{S}_{d'})) \right) = \mathbb{E}\left(R_d(\mathbf{S}_d, \mathbf{A}_d) + \gamma Q_\pi(\mathbf{S}_{d+1}, \pi(\mathbf{S}_{d+1})) \right) \tag{25}$$

where γ is the discount factor. The objective of the DRL agent is to find the optimal policy π^* that satisfies:

$$Q_{\pi^*}(\mathbf{S}_d, \mathbf{A}_d) = \mathbb{E}\left(R_d(\mathbf{S}_d, \mathbf{A}_d) + \gamma \max_{\mathbf{A}_{d+1}} Q_{\pi^*}(\mathbf{S}_{d+1}, \mathbf{A}_{d+1}) \right) \tag{26}$$

2) TD3 Algorithm

TD3 [23] is a popular DRL algorithm that is improved from the deep deterministic policy gradient (DDPG) algorithm [24]. The structure of the TD3 algorithm is shown in Fig. 3. Similar to the DDPG algorithm, the TD3 algorithm also has both actor and critic artificial neural networks. The actor network approximates the

agent's policy π, and the critic network approximates the action value function Q_π. In other words, the actor takes states as input and gives actions with the policy, while the critic network estimates how many rewards the agent can obtain in the future after taking these actions. This is why they are called the actor and critic, and the action values provided by critic are used to improve the actor's policy by the policy gradient algorithm. The corresponding target networks are used to improve the training performance. The agent's historical interactions with the environment are stored in the replay buffer so that they can be used to train the agent in the future. Full steps of TD3 are given as Algorithm 1.

Algorithm 1: The TD3 Algorithm

1. Create the replay buffer. Initialize the environment.
   Create the actor network π_θ, two critic networks Q_{ϕ_1} and Q_{ϕ_2}.
   Create the corresponding target network $\pi_{\theta_{\text{targ}}}$, $Q_{\phi_{\text{targ},1}}$ and $Q_{\phi_{\text{targ},2}}$.
2. Randomly initialize the actor network;s parameters θ, and the two critic networks' parameters ϕ_1, ϕ_2.
   Set corresponding target networks' parameters: $\theta_{\text{targ}} \leftarrow \theta$, $\phi_{\text{targ},1} \leftarrow \phi_1$, $\phi_{\text{targ},2} \leftarrow \phi_2$.
3. **Repeat** until N_{epi} episodes have been finished:
4. Observe state $\boldsymbol{S}_d$, choose action with noises
   $\boldsymbol{A}_d = \text{clip}(\pi_\theta(\boldsymbol{S}_d) + \varepsilon, \boldsymbol{A}_{\min}, \boldsymbol{A}_{\max})$
   $\varepsilon = (\boldsymbol{A}_{\max} - \boldsymbol{A}_{\min})\text{clip}(\mathcal{N}(0,\sigma), -\varepsilon_{\max}, \varepsilon_{\max})$
5. Execute $\boldsymbol{A}_d$ in the environment.
6. Observe next state $\boldsymbol{S}_{d+1}$, reward R_d and episode ending signal done (1 if the episode ends, or 0 otherwise).
7. Store $(\boldsymbol{S}_d, A_d, R_d, \boldsymbol{S}_{d+1}, done)$ into the replay buffer.
8. **If** $done = 1$ **then** reset the environment.
9. **If** N_{start} episodes have been finished **then:**
10. Randomly choose a batch of samples from the replay buffer as $B = \{(\boldsymbol{S}_d, \boldsymbol{A}_d, R_d, \boldsymbol{S}_{d+1}, done)\}$.
11. Compute target actions for each sample in B as
    $\boldsymbol{A}_{d+1}(\boldsymbol{S}_{d+1}) = \text{clip}(\pi_{\text{targ}}(\boldsymbol{A}_{d+1}) + \varepsilon_{\text{targ}}, \boldsymbol{A}_{\min}, \boldsymbol{A}_{\max})$
    $\varepsilon_{\text{targ}} = (\boldsymbol{A}_{\max} - \boldsymbol{A}_{\min})\text{clip}(\mathcal{N}(0,\sigma_{\text{targ}}), -\varepsilon_{\text{targ,max}}, \varepsilon_{\text{targ,max}})$
12. Compute target values as
    $y(R_d, \boldsymbol{S}_{d+1}, done) = R_d + \gamma(1 - done)\min_{i=1,2} Q_{\phi_{\text{targ},i}}(\boldsymbol{S}_{d+1}, \boldsymbol{A}_{d+1}(\boldsymbol{S}_{d+1}))$
13. Update critic networks by one step of gradient descent
    $\nabla_{\phi_i} \frac{1}{|B|}\sum_B \left[Q_{\phi_i}(\boldsymbol{S}_d, \boldsymbol{A}_d) - y(R_d, \boldsymbol{S}_{d+1}, d)\right]^2$ for $i = 1, 2$
14. **If** critics have been updated for N_{targ} times **then:**
15. Update actor network by one step of gradient ascent
    $\nabla_\theta \frac{1}{|B|}\sum_{\boldsymbol{S}_d \in B} Q_{\phi 1}(\boldsymbol{S}_d, \pi_\theta(\boldsymbol{S}_d))$
16. Update target networks
    $\phi_{\text{targ,i}} \leftarrow \rho\phi_{\text{targ},i} + (1-\rho)\phi_i$ for $i = 1, 2$
    $\theta_{\text{targ}} \leftarrow \rho\theta_{\text{targ}} + (1-\rho)\theta$
17. **End if**
18. **End if**
19. **End repeat**

C. Decision Timeline Transfer Learning

The SCTPGD problem is a typical sparse reward problem [21] in terms of the allowance shortage penalty, since the GenCo will only receive the penalty at the end of the annual compliance cycle if it does not buy allowances for its emissions. This means that only in 1/365 of the days where the GenCos has incentive to buy allowances for its emissions, which makes it difficult for the DRL agent to learn a good carbon trading strategy. To address this issue, a decision timeline transfer learning (DTTL) method is proposed as shown in Fig. 4.

The idea of DTTL is straightforward. The reward sparsity will be mitigated if there are less time steps before the end of the annual compliance cycle. When the SCTPGD problem is simplified into a decision timeline where decisions are made once a month, the agent will be incentivized to buy non-speculative allowances every 12 time steps. Consequently, the sparsity is greatly mitigated and the agent can learn the effect of the annual compliance cycle quickly. However, making decisions monthly will limit the flexibility and performance of the agent, so transfer learning is used to transfer the monthly decision agent to a weekly decision agent, and finally back to a daily decision agent that we need.

To make the transfer easier, the environments before and after the transfer should be as similar as possible. For all the monthly, weekly and daily decision environments, the state and decision spaces are nearly the same. For the monthly decision environment:

$$\boldsymbol{S}_m = \left(t'_m, \lambda_{\text{C}_0,m}, \overline{\lambda}_{\text{E},m}, \sigma_{\text{E},m}, E_{m-1}, V_{m-1}\right) \tag{27}$$

$$\boldsymbol{A}_m = (v_m, v_{\text{A},m}) \tag{28}$$

For the weekly decision environment:

$$\boldsymbol{S}_w = \left(t'_w, \lambda_{\text{C}_0,w}, \overline{\lambda}_{\text{E},w}, \sigma_{\text{E},w}, E_{w-1}, V_{w-1}\right) \tag{29}$$

$$\boldsymbol{A}_w = (v_w, v_{\text{A},w}) \tag{30}$$

For the daily decision environment:

$$\boldsymbol{S}_d = \left(t'_d, \lambda_{\text{C}_0,d}, \overline{\lambda}_{\text{E},d}, \sigma_{\text{E},d}, E_{d-1}, V_{d-1}\right) \tag{31}$$

$$\boldsymbol{A}_d = (v_d, v_{\text{A},d}) \tag{32}$$

where $m \in [1, M]$ is the index of months; $w \in [1, W]$ is the index of weeks; and $d \in [1, D]$ is the index of days. Most of the elements of the states and actions have the same meaning as introduced in Section III.B, except the first element in the states, i.e., the current time has been changed to a relative time in the compliance cycle as:

$$t'_m = \begin{cases} m/M, & \text{if } m < M \\ 2, & \text{if } m = M \end{cases} \tag{33}$$

$$t'_m = \begin{cases} w/W, & \text{if } w < W \\ 2, & \text{if } w = W \end{cases} \tag{34}$$

$$t'_d = \begin{cases} d/D, & \text{if } d < D \\ 2, & \text{if } d = D \end{cases} \tag{35}$$

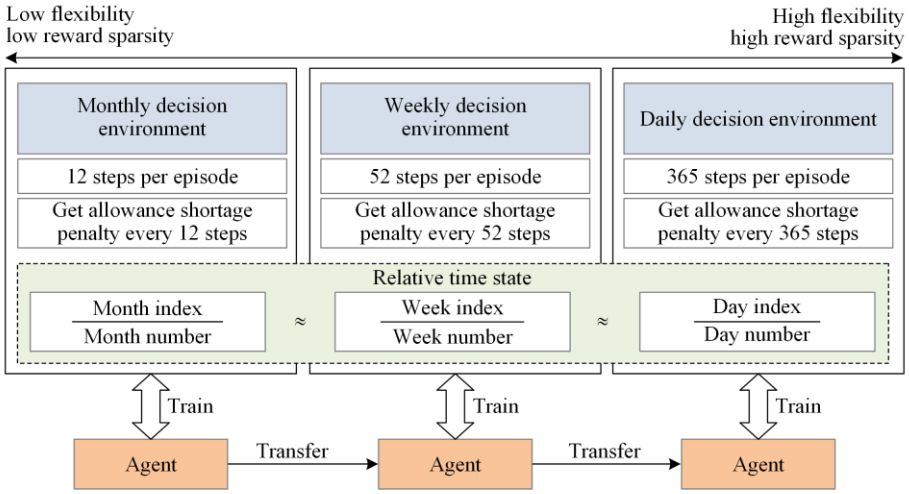


Fig. 4. Decision timeline transfer learning.

With the relative time state, the state variables for any given day remain essentially the same regardless whether the decision timeline is monthly, weekly or daily. This ensures that the agent trained in one decision timeline can easily adapt to another. It can also be noticed that when the time is at the end of the compliance cycle, the relative time state is set as 2, to help the agent better distinguish the end day and previous days.

Detailed steps of the DTTL method are provided as Algorithm 2. The replay buffer is cleared before training the agent on a different decision timeline because, although the state and action spaces remain largely the same, the state transitions differ across timelines due to the varying time intervals between consecutive timesteps. For lower-layer power generation decisions in the monthly (weekly) decision timelines, only one day in that month (week) is randomly chosen to represent all days in that month (week), so as to reduce computation burdens.

Algorithm 2: Decision Timeline Transfer Learning

1. Create the TD3 agent and the replay buffer.
2. Train the agent with the monthly decision environment using TD3 algorithm for N_{month} episode.
3. Empty the replay buffer.
4. Train the agent with the weekly decision environment using TD3 algorithm for N_{week} episode.
5. Empty the replay buffer.
6. Train the agent with the daily decision environment using TD3 algorithm for N_{day} episode.
7. **Done.**

D. Addressing the Asymmetry of the Carbon Trading Action Range

The TD3 algorithm adds noises to the actions to encourage explorations as shown in step 4 of Algorithm 1. However, this leads to issues when the daily carbon trading action range $[-v_{\text{sell}}, v_{\text{buy}}]$ is very asymmetric. For example, when speculation is discouraged by the GenCo or the government, the range may be set as $v_{\text{sell}} << v_{\text{buy}}$ or even $v_{\text{sell}} = 0$. Assuming the current daily carbon trading amount given by the actor network is $v_{d,0}$, the actual result after adding the exploring noise is:

$$v_d = \text{clip}(v_{d,0} + \varepsilon, -v_{\text{sell}}, v_{\text{buy}}) \tag{36}$$

where ε is a clipped zero-mean gaussian noise described in step 4 of Algorithm 1. Take the situation where $v_{\text{sell}} = 0$ for example, since the GenCos are allowed to buy all allowances for the whole year within a few days, $v_{d,0}$ is likely to be zero or a small value in most of the time. For $v_{\text{sell}} = 0$ and a small positive $v_{d,0}$, there is:

$$\mathbb{E}(v_d) = \mathbb{E}(\text{clip}(v_{d,0} + \varepsilon, 0, v_{\text{buy}})) > v_{d,0} \geqslant 0 \tag{37}$$

which means the exploring noises make v_d biased towards larger values. These biases will be accumulated and make the holding allowance V_d larger in the later days in the year, thereby preventing the agent from learning a good policy.

The proposed solution is to replace the asymmetric range with a symmetric one, and give a penalty to the carbon trading action outside the original range. Specifically, when $v_{\text{sell}} << v_{\text{buy}}$, the range $[-v_{\text{sell}}, v_{\text{buy}}]$ is replaced by $[-v_{\text{buy}}, v_{\text{buy}}]$, and an out-of-range penalty $\lambda_{\text{C},d}[-v_{\text{sell}} - v_d]^+$ is given to the agent.

Another issue is that the agent is not willing to approach the lower boundary $-v_{\text{sell}}$ when the out-of-range penalty is added, becoming a big problem when $v_{\text{sell}} = 0$. Since when the carbon price is high, the best strategy for the agent is to let $v_d = 0$. But when the agent is not willing to approach the lower boundary, it will always buy a small number of allowances even when the carbon price is high. This issue is caused by the severe nonlinearity of the reward function around the lower boundary, which is difficult for the neural network to approximate accurately. The solution proposed in this paper is to change the input structure of the critic network to help

the agent better distinguish the actions in and out of the range, so the agent is not afraid to approach the boundary. Specifically, v_d is split into v_d^+ and v_d^- before feeding into the critic networks, as:

$$v_d^+ = \max(v_d, -v_{\text{sell}}) \tag{38}$$

$$v_d^- = \min(v_d, -v_{\text{sell}}) \tag{39}$$

E. Risk Control by Allowance-Emission Difference Limit

Under the annual compliance cycle of the CTS, a GenCo's allowance holdings may differ from its cumulative emissions. This situation provides operational flexibility but also introduces financial risks. If the GenCo anticipates lower future carbon prices, it may postpone allowance purchases and increase power generation in expectation of reduced emission costs, potentially creating an allowance shortage before the expected price decline. However, if carbon prices do not fall as anticipated, the GenCo would face significant financial losses. In another situation where the GenCo believes that the carbon price will be high in the future and it buys many allowances now and wishes to sell them in the future to make money, so that an allowance surplus will exist before the high price comes. However, if the future carbon does not go high as expected, the GenCo will encounter a big loss. If the number of allowance shortage and surplus are limited across the year, the risks are also under control.

A risk control method by setting an allowance-emission difference limit (AEDL) for the SCTPGD problem is proposed. The objective of the risk control method is to ensure for all d, there is:

$$\left|V_d - E_d\right| \leqslant \Delta_{\text{AEDL}} \tag{40}$$

where Δ_{AEDL} is the value of the AEDL. In the training of the agent, constraint (40) is implemented by giving a penalty as $\lambda_{\text{pen}}\left[\left|V_d - E_d\right| - \Delta_{\text{AEDL}}\right]^+$ if (40) is violated.

After the DRL agent finishes its training, the day-by-day SCTPGD problem in the year can be solved by the HMDRL framework as shown in Fig. 5.

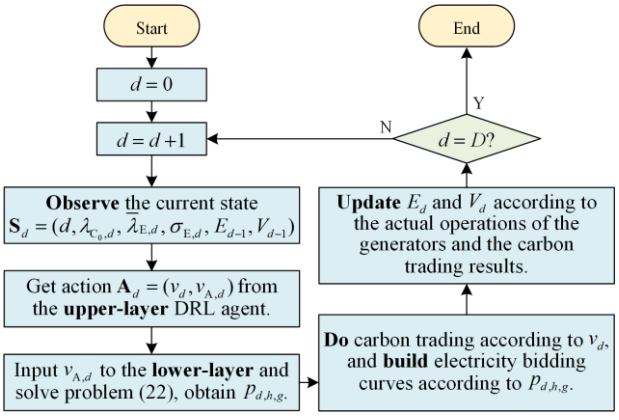


Fig. 5. Using the HMDRL framework to solve the day-by-day SCTPGD problem after the DRL agent finishes its training.

IV. Ex-post Speculative and Productive Trade Decomposition

When permitted to engage in carbon market speculation, the GenCo may actively trade allowances in response to price changes, causing its allowance holdings to fluctuate multiple times throughout the year. Meanwhile, the GenCo must purchase allowances to cover its emissions, which also alters its allowance inventory. It is an interesting problem to analyze that, after a year has ended, which part of the allowances are bought for speculation (speculative allowances) and which part are bought to cover its emissions from electricity production (productive allowances), so as to understand the behavior of the DRL agent. An two-step ex-post speculative and productive trade decomposition (SPTD) algorithm is proposed in this section to solve this problem as shown in Fig. 6.

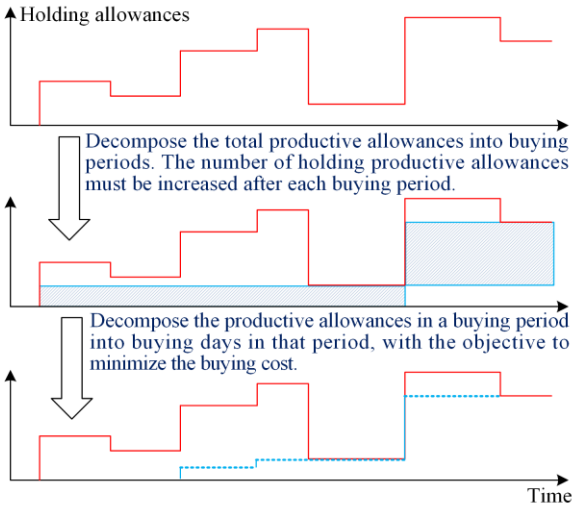


Fig. 6. The speculative and productive trade decomposition.

First, the total productive allowances are decomposed into buying periods. The number of holding productive allowances must be increased after each buying period, and the total number of productive allowances must be equal to the number of total emissions in the year. Specifically, with the historical v_d, V_d, $\lambda_{\text{C},d}$, search all $t_{\text{P},i}$ that satisfies:

$$V_{t_{\text{P},i}} < V_d \text{ for all } d > t_{\text{P},i} \tag{41}$$

$$t_{\text{P},i} < t_{\text{P},i+1} \tag{42}$$

where $t_{\text{P},i}$ is the end day of a buying period.

Second, the productive allowances within a buying period are decomposed into daily productive allowance buying amount $v_{\text{P},d}$, aiming to minimize the buying cost in the buying period:

$$\min_{v_{\text{P},d}} \lambda_{\text{C},d} v_{\text{P},d} \tag{43}$$

s.t.

$$d \in \left[1 + t_{P,i}, t_{P,i+1}\right] \tag{44}$$

$$V_{t_{P,i+1}} - V_{t_{P,i}} = \sum_d v_{\mathrm{P},d} \tag{45}$$

$$0 \leqslant v_{\mathrm{P},d} \leqslant [v_d]^+ \tag{46}$$

The above problem can be easily solved by iterations without a mathematical optimization solver.

It has to be clarified that, the result of the first step is objective, but the second step is subjective to the GenCo's preference. Since for every buying period, there must be a specific number of allowances being bought in that period and held to the end of the year. However, within a buying period, for instance, if the GenCo buys 10 units of allowances at 1 ¥/t in the first day, 10 units at 2 ¥/t on the second day, and then sells 10 units at 3 ¥/t on the third day, there is no way, from an ex-post perspective, to determine whether the sold allowances came from the first day, the second day, or a combination of both, since the resulting profit is the same in all cases. The algorithm proposed above is based on the assumption that the GenCo considers to have used a good strategy to purchase the productive allowances at the lowest cost.

V. Case Study

A. Case Settings

Numerical simulations are carried out to verify the effectiveness of the proposed methods. In the simulations, a GenCo with both gas-fired and coal-fired generators is considered, whose parameters are given in Appendix. The simulation program is written in Python language, the upper layer DRL agent is built with PyTorch, and the lower layer mathematical optimization problem is solved with Gurobi.

1) Electricity and Carbon Price Simulation Settings

The Ornstein-Uhlenbeck (OU) process is used to produce daily carbon price samples in the simulation environment, described by the stochastic differential equation as:

$$\mathrm{d}\lambda_t = \eta\left(\overline{\lambda} - \lambda_t\right)\mathrm{d}t + \sigma_{\mathrm{M}}\mathrm{d}B_t \tag{47}$$

where t is the continuous time variable; $\overline{\lambda}$, η and σ_{M} are parameters of the OU process; B_t is the standard Brownian motion. The OU process is a type of mean-reverting process that considers the price fluctuates stochastically but tends to go back to a specific mean value with a speed that is linear with the distance between them. It was used to model carbon and electricity prices in [25] and [26], respectively. The daily carbon prices of the EU ETS in the year of 2022 (given in Appendix) are used to estimate the parameters of the OU process of carbon price. (The EU ETS carbon prices are used because the carbon prices in China are relatively low, so that they have limited influence on the GenCo's behaviors. As the carbon prices in China are expected to rise in the future [27], the EU ETS prices are used to represent future cases). Free allocation is ignored as the EU ETS has stopped allocating free allowances to electricity generation sectors since 2013, to avoid providing windfall profits to GenCos [28]. It is also assumed that the carbon price is always in the range $\left[0, \lambda_{\mathrm{pen}}\right]$.

A linear model is used to reflect how the trading behavior affects the market price, as:

$$\lambda_{\mathrm{C},d} = \lambda_{\mathrm{C}_0,d} + k_{\mathrm{C}} v_d \tag{48}$$

where k_{C} is a positive coefficient, which means that the carbon price will rise when the GenCo buy allowances and drop when the GenCo sell allowances. The maximum daily buying number is set as $v_{\mathrm{buy}} = 10^6$ t. The carbon market response coefficient is set as $k_{\mathrm{C}} = 1/v_{\mathrm{buy}}$, which means that when the GenCo buys allowances with the number at the upper limit, the carbon price will rise to twice of the original one. Model (48) is based on the economic principle that the carbon price rises as the demand of allowances increases, which is similar to the stepped carbon price models in [11], [12]. In fact, the stepped carbon price models will be equivalent to model in (48) if their step numbers go to infinity.

The electricity price samples are produced as the sum of two parts. The first part is sampled from a given distribution and reflects that the electricity price is uncertain due to the stochasticity of power demand, renewable energy output, and actions of other GenCos. As shown in [29], [30], the emission costs of GenCos will be passed through to the electricity price, so the second part of the electricity price model is to reflect how the electricity price is affected by the carbon price.

$$\lambda_{\mathrm{E},d,h} = \lambda_{\mathrm{E}_0,d,h} + k_{\mathrm{EC},d}\lambda_{\mathrm{C}_0,d} \tag{49}$$

$$\lambda_{\mathrm{E}_0,d,h} \sim \mathrm{U}(\delta_{\min}\overline{\lambda}_{\mathrm{E},d,h}, \delta_{\max}\overline{\lambda}_{\mathrm{E},d,h}) \tag{50}$$

$$k_{\mathrm{EC},d} \sim \mathrm{U}(k_{\mathrm{EC,min}}, k_{\mathrm{EC,max}}) \tag{51}$$

where $\overline{\lambda}_{\mathrm{E},d,h}$ is the forecasted mean value of the electricity price; $k_{\mathrm{EC},d}$ is the correlation coefficient between the electricity and carbon prices, which is uniformly distributed in the range $\left[k_{\mathrm{EC,min}}, k_{\mathrm{EC,max}}\right]$; $\delta_{\min}$ and $\delta_{\max}$ are the lower and upper coefficients of the uniform distribution in (50), respectively. The spot market electricity prices of Guangdong, China in the year of 2022 (given in the appendix) are used as $\overline{\lambda}_{\mathrm{E},d,h}$ to generate stochastic electricity prices. Parameters related to the stochastic price models are provided in Appendix A.

2) DRL Settings

Critic networks in the DRL agent have 4 hidden layers, each with 256 neurons. The ReLU activation function is used for the hidden layers. Actor networks in the DRL agent also have 4 hidden layers with 256 neurons. The ReLU activation function is used for the hidden layers,

and the Tanh activation function is used for the output layer. The learning rate is set as 0.0001. Other DRL parameters are given in Table I, where both σ and $\varepsilon_{\max}$ have two values that are for the two action variables respectively. To keep the inputs to the neural networks not too large, prices are scaled by 10^{-3}, allowances and emissions are scaled by 10^{-6}, whereas other related values are scaled accordingly, e.g., carbon buying cost and electricity selling profit are scaled by 10^{-9}.

TABLE I
DRL PARAMETERS

Parameter	Value	Parameter	Value	Parameter	Value
γ	1	N_{start}	100	σ	0.05, 0.1
ρ	0.99	N_{month}	1000	$\varepsilon_{\max}$	0.1, 0.5
$\lvert B\rvert$	512	N_{week}	1000	σ_{targ}	0.0001
N_{targ}	2	N_{day}	4000	$\varepsilon_{\text{targ,max}}$	0.0002

B. Effect of Annual Compliance Cycle and Market Response

To validate the necessity of incorporating both the annual compliance cycle and market response in the SCTPGD problem, and to evaluate the effectiveness of the HMDRL framework, 5 cases are studied with different combinations of environments and decision methods. To focus on evaluating the synergistic performance of carbon trading and generation decision, speculation is disallowed by setting $v_{\text{sell}} = 0$, and the analysis of carbon market speculation is provided in later sections.

Two decision methods are involved in the 5 cases.

HMDRL: The hybrid mathematical-DRL optimization framework proposed in this paper.

Myopic: This is the most popular strategy currently used in the literature [3]–[13]. The annual compliance cycle of the CTS is ignored, and the GenCo has to buy allowances at the same time that emissions happen. In this method, the GenCo directly solves the one-day SCTPGD problem with the mathematical solver.

The settings of the 5 cases are as follows.

Case 1: The Myopic method is performed in an environment that the carbon price is not affected by the GenCo's action, i.e., $k_{\text{C}} = 0$, and the method is aware of this.

Case 2: The HMDRL method is performed in an environment that the carbon price is not affected by the GenCo's action, i.e., $k_{\text{C}} = 0$. The DRL agent is also trained with $k_{\text{C}} = 0$.

Case 3: The Myopic method is performed in an environment that the carbon price is affected by the GenCo's action, i.e., $k_{\text{C}} = 1/v_{\text{buy}}$, and the method is aware of this.

Case 4: The HMDRL method is performed in an environment that the carbon price is affected by the GenCo's action, i.e., $k_{\text{C}} = 1/v_{\text{buy}}$, but the GenCo ignores this, and the DRL is trained with $k_{\text{C}} = 0$.

Case 5: The HMDRL method is performed in an environment that the carbon price is affected by the GenCo's action, i.e., $k_{\text{C}} = 1/v_{\text{buy}}$, and the GenCo takes the market response into account in the training of the DRL agent.

With each one of the 5 cases, 365 day-by-day rolling decisions are performed in 1000 stochastically sampled scenarios, and the results are given in Table II.

TABLE II
AVERAGE RESULTS OF 1000 SCENARIOS IN THE 5 CASES

Case number	Consider annual compliance cycle	Market has response	Consider market response	Profit (¥)	Average allowance buying price (¥/t)	Average electricity selling price (¥/MWh)	Emission (t)	Electricity generation (MWh)	Maximum daily allowance buying amount (t)
1	No	No	No	3.30×10^{8}	593.31	786.97	2 332 732.02	1 298 810.76	8910.63
2	Yes	No	No	4.07×10^{8}	528.82	736.77	3 306 634.15	1 924 932.16	432 585.57
3	No	Yes	Yes	3.24×10^{8}	597.77	789.42	2 286 283.13	1 264 901.11	8910.63
4	Yes	Yes	No	1.68×10^{8}	653.02	736.77	3 306 634.15	1 924 932.16	432 585.57
5	Yes	Yes	Yes	3.64×10^{8}	558.97	751.95	2 991 101.61	1 730 661.73	64 944.29

First, the importance of the annual compliance cycle is analyzed. By comparing the results of Case 1 and 2, it can be seen that the HMDRL method obtains a higher profit than the Myopic method. Meanwhile, the average allowance buying price of HMDR method is lower than the Myopic method, indicating that the GenCo can buy allowances with a lower cost. As a result, the average electricity selling price of the HMDRL is also lower than the Myopic method, due to the GenCo's ability to make profits under lower electricity prices with the lower emission cost obtained by the HMDRL method. The reason why HMDRL outperforms the Myopic method can be explained by Fig. 7 (Scenario 2 is taken as the representative example, and more scenarios are given in the supplemental materials.). Since the HMDRL method takes the advantage of the annual compliance cycle, it can buy allowances when the price is low instead of when the emissions happen. However,

the Myopic has to buy allowances at the same time as the emissions, without utilizing carbon price fluctuations.

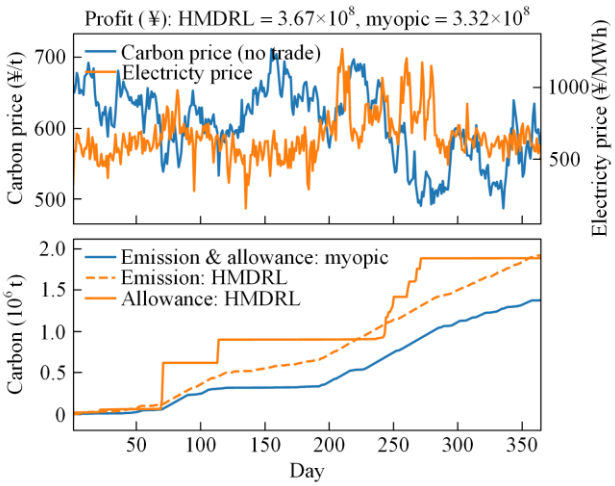


Fig. 7. Comparison of the HMDRL and Myopic method.

The influence of carbon market responses can be observed in the results of Case 3, 4 and 5. It can be seen in Table II that, when the GenCo's action influences the carbon price, as in Case 3 and 5, the HMDRL method can still obtain a profit that is significantly higher than the Myopic method if the agent takes the market response into account. The reason is the same as in the analysis of Case 1 and 2. But if the agent ignores the carbon market response as in Case 4, the utilization of the annual compliance cycle leads to a deteriorated result with the lowest profit among all the 5 cases. It can be seen in Table II that the average allowance buying price in Case 4 is the highest but the average electricity selling price is still as low as in Case 2. Because the agent considers that it can buy allowances at a low price, it is willing to sell electricity at low prices. However, as shown in Fig. 8 (Scenario 2 is taken as the representative example, and more scenarios are given in the supplemental materials.), the market price rise significantly for the large demand, leading to the agent to buy allowances at prices higher than it expected. Profits are lost because of the higher emission cost and the lower electricity overselling prices.

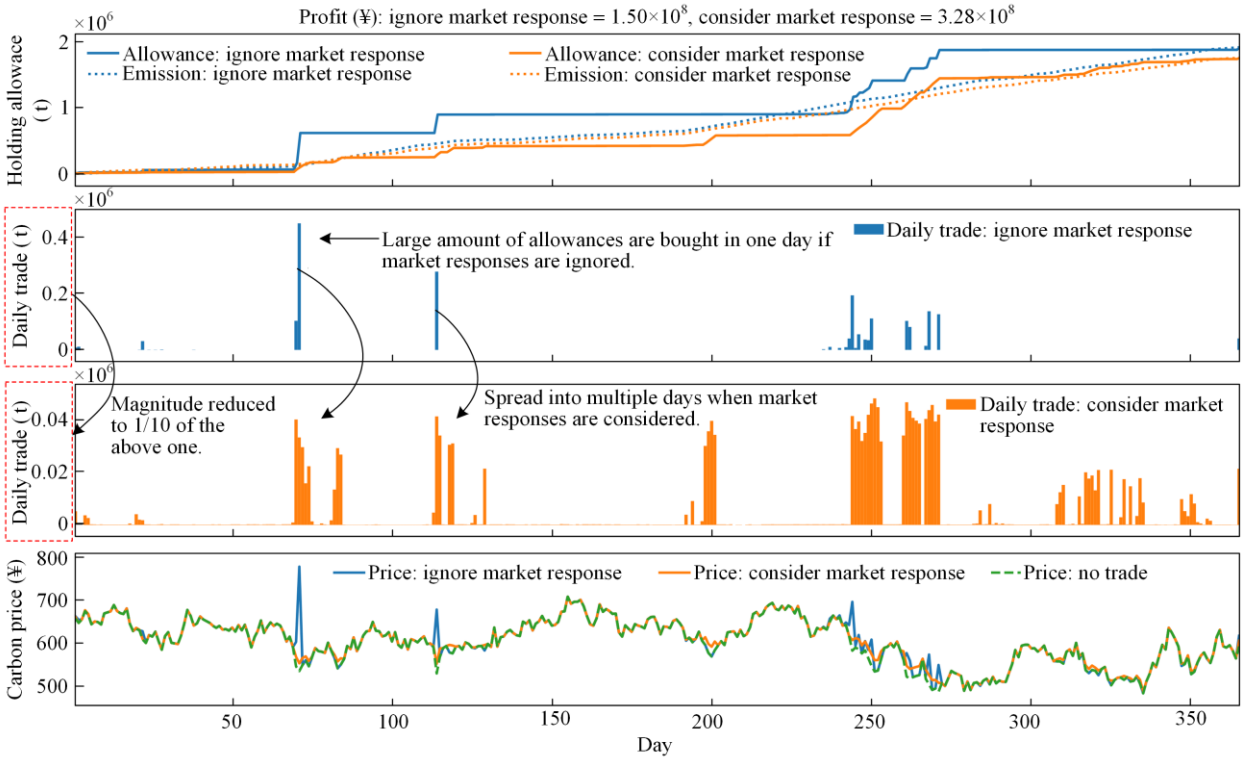


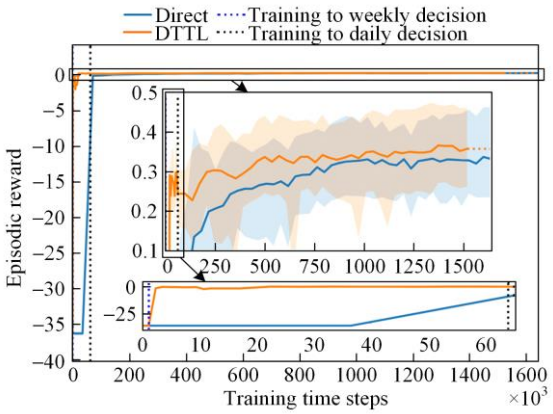
Fig. 8. Example of influence of the carbon market response.

Figure 8 demonstrates how HMDRL handles the carbon market response effectively. It can be seen that HMDRL adapts well to the environment with market response. It tries to buy allowances when the price is low, whereas unlike the case without market response, the buying trades are spread over many low-price days, rather than concentrated in a small number of days. This can also be observed in Table II that the maximum daily allowance buying amount in Case 5 is much smaller than Case 4. In such a way, the trading amount in a single day is limited and the market price does not rise significantly, so that the GenCo can still lower its emission cost by utilizing the annual compliance cycle.

Meanwhile, it is interesting to noticed that, in Table II the yearly total emissions in Case 5 are less than Case 2. The same situation exists between Case 3 and Case 1. This shows that the market response of the CTS can help to reduce emissions from the GenCos.

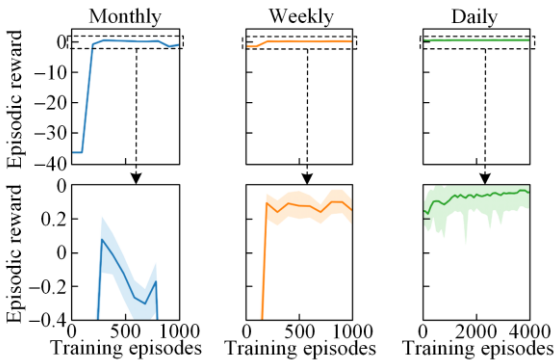
C. Verification of Decision Timeline Transfer Learning

As discussed in Section III.C, the SCTPGD problem is a sparse reward problem in terms of allowance shortage penalty, and therefore, the DTTL method is proposed to address this issue. To verify its effectiveness, the proposed method is compared to that training the agent directly on the daily decision environment. Specifically, with the DTTL method, the agent is trained on the monthly, weekly and daily decision environment for 1000, 1000 and 4000 episodes respectively, which has 1 524 000 time steps in total. For the direct training method, the agent is trained on the daily decision environment for 4500 episodes, accounting for 1 642 500 time steps in total. The number of time steps must be considered because the total training time grows linearly with it. This is due to the fact that each time step across the three decision timelines incurs similar computational cost, invoking Gurobi to solve the lower-layer problem and train the agent. Carbon market responses are considered while allowances selling is not allowed in the test. For every 100 training episodes, the agent is tested on the daily decision environment for 100 stochastic sampled scenarios. Results are shown in Fig. 9 and Fig. 10.



Note: The lower row is the enlargement of the upper row. Shadows are the distribution of the 100 scenarios.

Fig. 9. Learning curves of direct learning and DTTL.



Note: In the three columns, the agent is being trained in the monthly, weekly, daily decision environment respectively, and tested in the daily decision environment. The second row is the enlargement of the first row. Shadows are the distribution of the 100 scenarios.

Fig. 10. Episodic reward in the DTTL process.

As shown in Fig. 9, with the DTTL method, the agent learns significantly faster than with direct learning and consistently outperforms it throughout the entire training process. This is because monthly and weekly decision episodes contain only 12 and 52 time steps respectively, allowing the agent to receive end-of-compliance-cycle feedback more frequently than in the daily decision environment. Figure 10 shows that the agent learns very fast in the monthly and weekly decision environment, and performs well after been transferred to another decision timeline. This verifies that the proposed relative time state makes the tasks in different decision timelines similar to the agent, which helps the agent to efficiently adapt to the new environment.

D. Effectiveness of the Measures for Asymmetric Carbon Trading Range

As introduced in Section III.D. The exploring noises in the TD3 algorithm can bias the agent's action if the carbon trading range $[-v_{\text{sell}}, v_{\text{buy}}]$ is too asymmetric as $v_{\text{sell}} << v_{\text{buy}}$. To solve this problem, it is proposed in Section III.D that a symmetric carbon trading range is used and a penalty is applied to the agent when the carbon trading action is out of the original range. Meanwhile, the trading action is split into two variables before feeding into the critic network to address the issue caused by the penalty. These measures are tested with $v_{\text{sell}} = 0$, while all agents are fully trained with the number of episodes described above.

As shown in Fig. 11 (Scenario 2 is taken as the representative example, and more scenarios are given in the supplemental materials.), with the asymmetric carbon trading range $[0, v_{\text{buy}}]$, the TD3 algorithm always adds a positive noise to the carbon trading action. Since this occurs every day throughout the year, the positive noises in the exploring process are accumulated, resulting in an excessively large number of holding allowances. Most of the agent's experiences fall into such situations, which prevent it from learning an effective policy. In contrast, when the carbon trading range is replaced with a symmetric one, the noises becomes unbiased, and the number of holding allowances fluctuates around the normal levels, which facilitates more effective exploration by the agent.

It can be seen in Table III that by splitting the input of carbon trading action to the critic network, higher profits and lower allowance buying prices are obtained. This can be explained in Fig. 12 (Scenario 5 is taken as the representative example, and more scenarios are given in the supplemental materials.) that, after the asymmetric range being replaced by the symmetric one, the out-of-range penalty prevents the agent from approaching the original boundary. Consequently, in the case of Fig. 12, when the optimal action is to buy zero allowance, the agent will still output a small positive value, leading to a suboptimal policy. Splitting the carbon trading action at the lower boundary helps the agent better distinguish the actions in and out of the range, reducing its hesitation to approach the boundary.

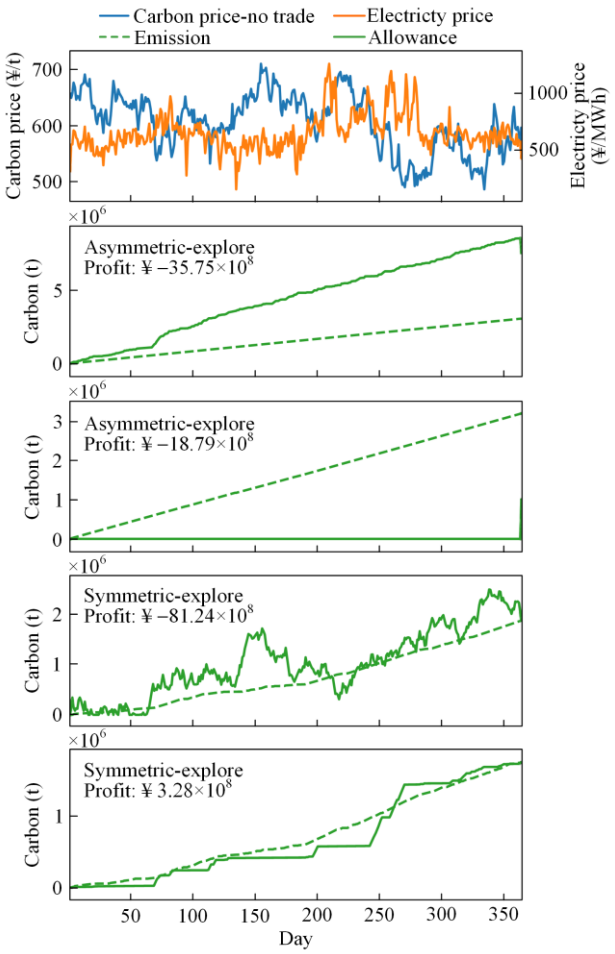


Fig. 11. Effectiveness of using a symmetric carbon trading range.

TABLE III
AVERAGE RESULTS OF 1000 SCENARIOS WITH SINGLE AND SPLIT INPUTS

Carbon trading action input to critic network	Profit (¥)	Average allowance buying price (¥/t)
Single input	3.42×10^8	568.2454
Split inputs	3.64×10^8	558.9693

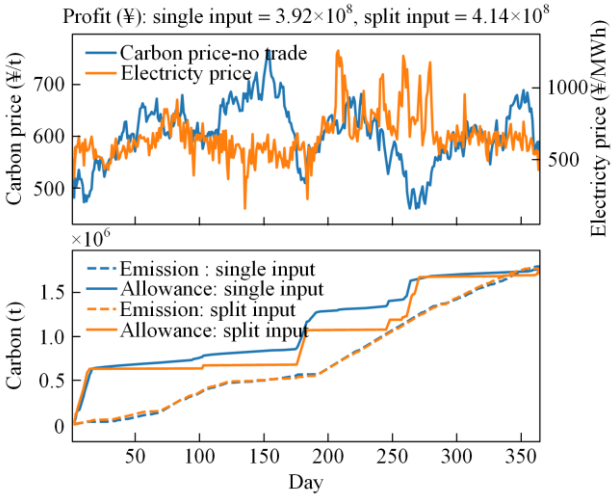


Fig. 12. Effectiveness of using split inputs to critic network.

E. Verification of AEDL Risk Control

The risk control method by setting the allowance-emission difference limit (AEDL) is verified in this section. In the simulation, the carbon market response is considered.

Figure 13 shows the mean value and standard deviation of the profits in 1000 scenarios with different AEDLs, with and without allowance selling respectively. When AEDL is 0, the Myopic method is used. It can be seen that both the means and standard deviations with every AEDL are larger when selling is allowed, as the GenCo can get more profit from speculations in the carbon market while taking more risks. As AEDL becomes larger, the GenCos makes more profits in general both with and without allowance selling, since larger AEDL gives it more flexibility. When selling is allowed, the standard deviation becomes significantly higher with large AEDL, which shows that setting AEDL can control the risk for the GenCo. When selling is not allowed, the standard deviation remains relatively small with all AEDLs, since the GenCo does not buy allowances more than it needs, and the market response makes it spread the buying actions over multiple day as discussed in Fig. 8.

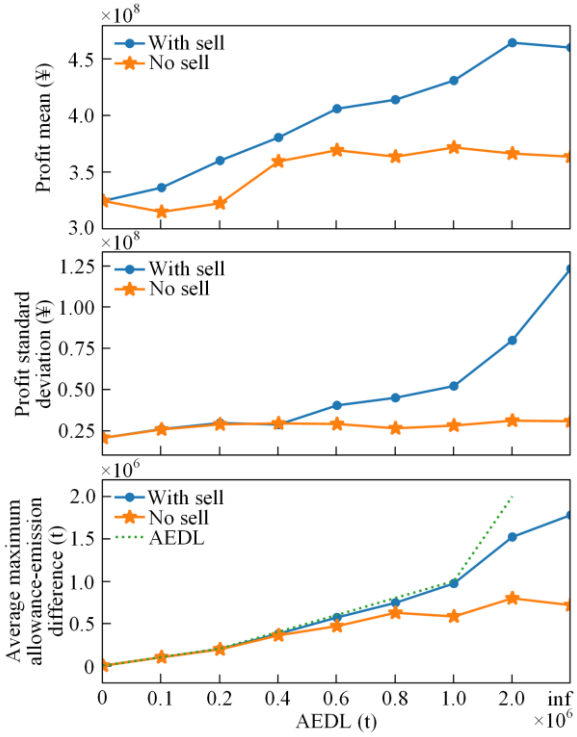


Fig. 13. Mean value and standard deviation of the profits in 1000 scenarios with different AEDLs.

Some phenomena in Fig. 13 may appear counterintuitive, for example, the mean profit does not always increase with a larger AEDL, but they can be reasonably explained. The first category is that when AEDL changes from 0 to 0.1×10^6 the profit drops without allowances selling. This is because the DRL agent only observe the

mean and standard deviation of the electricity prices rather than using every hourly electricity price to make allowance usage decisions as the Myopic method does, so the flexibility provided by the small AEDL cannot compensate this disadvantage. The second category includes when AEDL is changed from 2×10^6 t to infinite with allowance selling, and when AEDL is larger than 0.6×10^6 t without allowance selling. These can be explained by the third row in Fig. 13, that the average maximum allowance-emission difference starts to be smaller than AEDL when AEDL is large enough, which means that the GenCo's actions are not limited by AEDL in more and more scenarios as AEDL increases. Thus, once AEDL is sufficiently large, further increases provides little additional flexibility to the GenCo, leaving little room for further profit improvement. Under such conditions, the optimal profits achieved by HMDRL become similar across different AEDLs, and small deviations from the optimal policy can disrupt the monotonic trend observed in the first two rows of Fig. 13. This is not an issue, however, because all the resulting profits remain close to their optimal values.

Figure 14 (Scenario 5 is taken as the representative example, and more scenarios are given in the supplemental materials.) shows an example to illustrate how AEDL affects GenCo's actions.

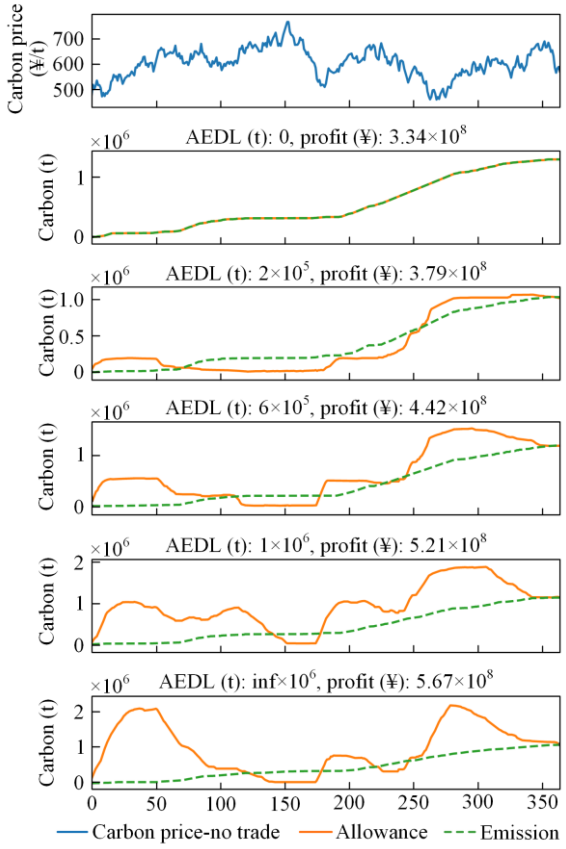


Fig. 14. Effectiveness of AEDL.

It can be found that with small AEDL, the curve of holding allowances is close to the curve of emissions, so that there are small number of allowances that must be bought or sold in the future, and thus the risk is controlled.

F. Speculation Analysis

The main objective for studying the speculation is to analyze its influences. The ex-post SPTD algorithm is applied to decompose the allowances trading history into speculative parts and productive parts. An example is shown in Fig. 15 (Scenario 5 is taken as the representative example, and more scenarios are given in the supplemental materials.). It is seen that the algorithm successfully decomposes the allowances trading curve into speculative and productive allowances as expected. Such results can be used to analyze the behaviors of the GenCo. An example application is provided in Table IV, where the average allowance trading price is calculated as $\sum\lambda_{\mathrm{C},d}v_d/\sum v_d$, the allowance buying price is calculated as $\sum\lambda_{\mathrm{C},d}[v_d]^+/\sum[v_d]^+$, and the productive allowance buying price is calculated as $\sum\lambda_{\mathrm{C},d}v_{\mathrm{P},d}/\sum v_{\mathrm{P},d}$.

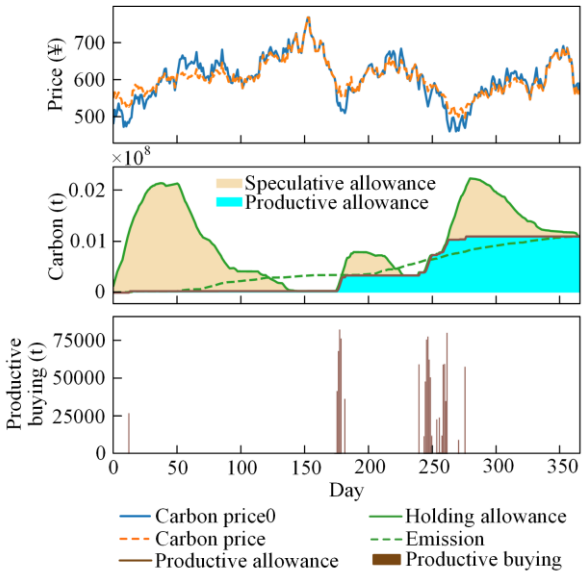


Fig. 15. Results of the SPTD algorithm.

It can be found in Table IV that, when the GenCo is allowed to speculate in the carbon market, it obtains higher profit than without speculation. However, even the average allowance trading price and the allowance buying price are lower when the GenCo speculates, the GenCo generates less electricity than without speculation, which seems to be unreasonable. With the ex-post SPTD algorithm, it can be seen that the productive allowance buying price is higher when it speculates, which explains the previous phenomenon. Because the productive allowance buying price reflects the actual cost of the allowances used to cover emissions, whereas the other two prices do not. This means that speculation can increase the emission cost of the GenCo, and consequently, the GenCo reduces its emissions. It should be

noticed that the productive allowance buying price calculated by the ex-post SPTD algorithm may be underestimated since cost minimization is used to decompose the productive allowances into days.

TABLE IV
AVERAGE VALUES OF 1000 SCENARIOS WITH AND WITHOUT SPECULATION

Speculation	Profit (¥)	Speculative profit (¥)	Productive profit (¥)	Emission (t)	Electricity generation (MWh)	Average allowance trading price (¥/t)	Average allowance buying price (¥/t)	Average Productive allowance buying price (¥/t)	Average electricity selling price (¥/MWh)
No	3.64×10^8		3.64×10^8	1 730 661.73	2 991 101.61	558.97	558.97	558.97	751.95
Yes	4.61×10^8	1.59×10^8	3.02×10^8	1 083 023.77	2 011 797.16	436.52	535.52	583.83	792.00

Table V shows how the GenCo's actions affect the carbon market prices. In the first line, the GenCo does not trade allowance at all. In the second line, the GenCo only buys allowances without selling. In the third line, the GenCo buys and sells allowances for production and speculation. It can be found that the average prices become slightly higher in the second and third lines since the GenCo needs to buy allowances for its emissions, thereby increasing the demand in the market. It can also be found that the trading actions can mitigate the carbon price fluctuations as the standard deviation becomes lower when the GenCo participates in the market. For the non-speculative one, the minimum prices are raised since the GenCo tends to buy allowances when the price is low. For the speculative one, both the minimum and maximum prices are pushed towards the average value, since the GenCo tries to buy allowances when the price is low and sell them when the price is high, as shown in Fig. 15.

TABLE V
CARBON PRICE STATISTICS OF 1000 SCENARIOS

GenCo action	Yearly average (¥/t)	Yearly standard deviation (¥/t)	Yearly maximum (¥/t)	Yearly minimum (¥/t)
No trade	603.11	52.64	733.51	473.01
No speculation	605.66	50.04	735.01	489.10
With speculation	603.95	44.29	723.83	500.16

In summary, under the setting of this paper, speculations of the GenCo help to increase the profit of the GenCo, stabilize carbon price and reduce emissions of the GenCo, which are beneficial for both the GenCo and the society.

VI. CONCLUSION

In this paper, both annual compliance cycle and market response of the carbon trading system are considered in the SCTPGD problem, and a HMDRL optimization framework that combines the DRL and mathematical methods is used to solve the problem. Case studies show that:

1) The HMDRL optimization framework leverages the annual compliance cycle to reduce the GenCo's emission costs and enhance profitability. By considering the carbon market response effect, HMDRL helps the GenCo spread its trades across multiple days to avoid severe impacts on the carbon market price.

2) The DTTL method can mitigate the sparsity of allowance shortage penalty, and help the agent learn faster and better than the direct learning approach.

3) The ex-post SPTD algorithm can decompose the allowances trading history into speculative parts and productive parts to analyze the GenCo's behaviors. Under the setting of this paper, speculations in the carbon market help to increase the profit of the GenCo, stabilize carbon market price and reduce emissions of the GenCo, which are beneficial to both the GenCo and the society.

Several issues remain for future studies. For instance, the bidding and clearing process of the electricity and carbon markets can be considered, and the decision problem of GenCo's with both fossil and renewable energy can also be studied.

APPENDIX A

TABLE AI
PRICE MODEL PARAMETERS

$\bar{\lambda}$	σ_M	η	δ_{min}	δ_{max}	$k_{EC,min}$	$k_{EC,max}$	$\bar{\lambda}_{E,d,h}$
602.78	263.60	0.041	0.9	1.1	0.1	0.3	Fig. A1

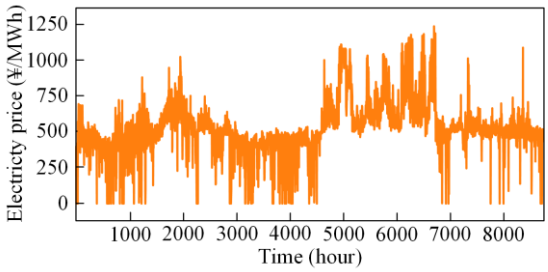
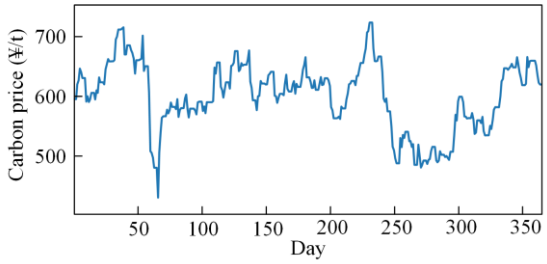


Fig. A1. Hourly electricity prices of Guangdong, China, 2022.



Note: Converted from EUR to Chinese Yuan at a ratio of 1:7.39. Missing values are filled the nearest data.

Fig. A2. Carbon prices of the EU ETS in the year of 2022.

TABLE A II
GENERATOR PARAMETERS

Index	Fuel	$P_{\max,g}$ (MW)	$P_{\min,g}$ (MW)	$P_{U,g}$ (MW)	$P_{D,g}$ (MW)	$h_{U,g}$ (hour)	$h_{D,g}$ (hour)	$\alpha_{g,1}$ (t/MWh)	$\alpha_{g,0}$ (t/h)	k_g (t/t)	$\lambda_{W,g}$ (¥/t)	$\lambda_{U,g}$ (¥)	$\lambda_{D,g}$ (¥)
1	Coal	320	120	120	120	4	4	0.3132	11.20	2.26	600	800 000	180 000
2	Gas	300	100	600	600	1	1	0.1086	6.12	3.08	3000	100 000	100 000

ACKNOWLEDGMENT

Not applicable.

AUTHORS' CONTRIBUTIONS

Shouyuan Shi: methodology, formal analysis, writing–original draft. Zhenning Pan: writing–review & editing, Funding acquisition. Junbin Chen: writing–review & editing. Tao Yu: supervision, funding acquisition. All authors read and approved the final manuscript.

FUNDING

This work is jointly supported by the Natural Science Foundation of China-Smart Grid Joint Fund of State Grid Corporation of China (No. U2066212); the National Natural Science Foundation of China (No. 52207105); and the Key Science and Technology Projects of China Southern Power Grid Corporation (No. 066600KK52222023).

AVAILABILITY OF DATA AND MATERIALS

More simulation results are available at https://github.com/shishouyuan/PaperData_PCMP_HMDRL_Electricity_Carbon_Trading.

DECLARATIONS

Competing interests: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

AUTHORS' INFORMATION

Shouyuan Shi received the B.Eng. degree in electrical engineering from Shandong University, Jinan, China, in 2017. He is currently pursuing the Ph.D. degree with South China University of Technology. His major research interests include optimization problem in the electricity and carbon markets, and management of distributed energy resources in smart power grid.

Zhenning Pan received the B.Eng. and Ph.D. degrees in electrical engineering from South China University of Technology, Guangzhou, China, in 2016 and 2021, respectively. His major research interests include intelligent operation and optimization of smart grid, and demand response.

Junbin Chen received the M.Eng. degree in electrical engineer from South China University of Technology, Guangzhou, China, in 2020, where, he is currently pursuing the Ph.D. degree with the College of Electric Power. His main research interests include reinforcement learning and machine learning for smart power grid.

Tao Yu received the B.Eng. degree in electrical power system from Zhejiang University, Hangzhou, China, in 1996, and the Ph.D. degree in electrical engineering from Tsinghua University, Beijing, China, in 2003. He is currently a Professor with the College of Electric Power, South China University of Technology, Guangzhou, China. He is also with Guangzhou Provincial Key Laboratory of Intelligent Measurement and Advanced Metering of Power Grid, Guangzhou, China. His special fields of interest include nonlinear and coordinated control theory, artificial intelligence techniques in planning, and operation of power systems.

REFERENCES

[1] European Commission, (2024, Jan.) "Monitoring, reporting and verification of EU ETS emissions," European Commission, [Online]. Available:https://climate.ec.europa.eu/eu-action/eu-emissions-trading-system-eu-ets/monitoring-reporting-and-verification-eu-ets-emissions_en

[2] Ministry of Ecology and Environment of the People's Republic of China, (2024, Jan.) "Carbon emission trading management measures (Trial)." [Online]. Available: https://www.gov.cn/zhengce/zhengceku/2021-01/06/content_5577360.htm

[3] Q. Tan, Y. Ding, and Q. Ye *et al*., "Optimization and evaluation of a dispatch model for an integrated wind-photovoltaic-thermal power system based on dynamic carbon emissions trading," *Applied Energy*, vol. 253, Nov. 2019.

[4] P. Mathuria, R. Bhakar, and F. Li, "GenCo's optimal power portfolio selection under emission price risk," *Electric Power System Research*, vol. 121, pp. 279-286, Apr. 2015.

[5] Y. Xiang, M. Fang, and J. Liu *et al*., "Distributed dispatch of multiple energy systems considering carbon trading," *CSEE Journal of Power and Energy Systems*, vol. 9, no. 2, pp. 459-469, Mar. 2023.

[6] P. Mathuria and R. Bhakar, "GenCo's integrated trading decision making to manage multimarket uncertainties," *IEEE Transactions on Power Systems*, vol. 30, no. 3, pp. 1465-1474, May 2015.

[7] Q. Chen, C. Kang, and Q. Xia *et al*., "Optimal flexible operation of a co₂ capture power plant in a combined energy and carbon emission market," *IEEE Transactions on Power Systems*, vol. 27, no. 3, pp. 1602-1609, Aug. 2012.

[8] T. Liang, L. Chai, and J. Tan *et al*., "Dynamic optimization of an integrated energy system with carbon capture and power-to-gas interconnection: a deep reinforcement learning-based scheduling strategy," *Applied Energy*, vol. 367, Aug. 2024.

[9] Y. Zhang, Z. Mei, and X. Wu *et al*., "Two-step diffusion policy deep reinforcement learning method for low-carbon multi-energy microgrid energy management," *IEEE Transactions on Smart Grid*, vol. 15, no. 5, pp. 4576-4588, Sep. 2024.

[10] X. Wei, Y. Xu, and H. Sun *et al*., "Day-ahead optimal dispatch of a virtual power plant in the joint energy-reserve-carbon market," *Applied Energy*, vol. 356, Feb. 2024.

[11] C. Wei, Y. Wang, and Z. Shen *et al*., "AUQ–ADMM algorithm-based peer-to-peer trading strategy in large-scale interconnected microgrid systems considering carbon trading," *IEEE Systems Journal*, vol. 17, no. 4, pp. 6248-6259, Dec. 2023.

[12] X. Zhang, X. Guo, and X. Zhang, "Bidding modes for renewable energy considering electricity-carbon integrated market mechanism based on multi-agent hybrid game," *Energy*, vol. 263, Jan. 2023.

[13] K. Jiang, N. Liu, and X. Yan, "Modeling strategic behaviors for GenCo with joint consideration on electricity and carbon markets," *IEEE Transactions on Power Systems*, vol. 38, no. 5, pp. 4724-4738, Sep. 2023.

[14] S. J. Kazempour, M. P. Moghaddam, and M. R. Haghifam *et al*., "Dynamic self-scheduling of a fuel and emission constrained power producer under uncertainties," in *2009 IEEE/PES Power Systems Conference and Exposition*, Seattle, USA, Mar. 2009, pp. 1-10.

[15] S. Deng, H. Chen, and D. Xiao *et al*., "A joint decision making model for power generators to participate in the carbon market and the medium-and long-term power markets," *Power System Protection and Control*, vol. 50, no. 22, pp. 1-10, Nov. 2022. (in Chinese)

[16] X. Li, C. Yu, and Z. Xu *et al*., "A multimarket decision-making framework for GENCO considering emission trading scheme," *IEEE Transactions on Power Systems*, vol. 28, no. 4, pp. 4099-4108, Nov. 2013.

[17] Y. Liu, H. Sun, and B. Meng *et al*., "How to purchase carbon emission right optimally for energy-consuming enterprises? analysis based on optimal stopping model," *Energy Economics*, vol. 124, Aug. 2023.

[18] S. Shi, T. Yu, and C. Lan *et al*., "Estimating the actual emission cost in an annual compliance cycle: synergistic generation and carbon trading optimization for price-taking generation companies," *Applied Energy*, vol. 376, Dec. 2024,

[19] F. Beltrán, W. de Oliveira, and E. C. Finardi, "Application of scenario tree reduction via quadratic process to medium-term hydrothermal scheduling problem," *IEEE Transactions on Power Systems*, vol. 32, no. 6, pp. 4351-4361, Nov. 2017.

[20] R. Zhu, "DRL based low carbon economic dispatch by considering power transmission safety limitations in internet of energy," *Internet of Things*, vol. 24, Dec. 2023.

[21] S. Li, X. Wang, and W. Zhang *et al*., "A model-based approach to solve the sparse reward problem," in *2021 4th International Conference on Pattern Recognition and Artificial Intelligence (PRAI)*, Aug. 2021, pp. 476-480.

[22] A. J. Conejo, F. J. Nogales, and J. M. Arroyo, "Price-taker bidding strategy under price uncertainty," *IEEE Transactions on Power Systems*, vol. 17, no. 4, pp. 1081-1088, Nov. 2002.

[23] S. Fujimoto, H. van Hoof, and D. Meger (2024, Jun.), "Addressing function approximation error in actor-critic methods," [Online]. Available: http://arxiv.org/abs/1802.09477.

[24] T. P. Lillicrap (2024, Aug.), "Continuous control with deep reinforcement learning," [Online]. Available: http://arxiv.org/abs/1509.02971

[25] L. Wang, "Two-stage stochastic planning for integrated energy systems accounting for carbon trading price uncertainty," *International Journal of Electrical Power & Energy Systems*, vol. 143, Dec. 2022.

[26] J. Kettunen, A. Salo, and D. W. Bunn, "Optimization of electricity retailer's contract portfolio subject to risk preferences," *IEEE Transactions on Power Systems*, vol. 25, no. 1, pp. 117-128, Feb. 2010.

[27] S. Qi, S. Cheng, and X. Tan *et al*., "Predicting China's carbon price based on a multi-scale integrated model," *Applied Energy*, vol. 324, Oct. 2022.

[28] European Commission (2024, May), "Allocation to modernise the energy sector," [Online]. Available: https://climate.ec.europa.eu/eu-action/eu-emissions-trading-system-eu-ets/free-allocation/allocation-modernise-energy-sector_en

[29] H. Wang, T. Feng, and C. Zhong, "Effectiveness of CO_2 cost pass-through to electricity prices under 'electricity-carbon' market coupling in China," *Energy*, vol. 266, Mar. 2023.

[30] W. Kim, D. Chattopadhyay, and J. Park, "Impact of carbon cost on wholesale electricity price: a note on price pass-through issues," *Energy*, vol. 35, no. 8, pp. 3441-3448, Aug. 2010.