Proceedings of the 3rd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2025.26411

Dynamic Hedging Strategies for Carbon Asset Pricing with Deep Reinforcement Learning

Suwan Hu

Monash University, Melbourne, Australia
suwhu0030@163.com

Abstract. Carbon allowance prices are buffeted by regime shifts from policy announcements, weather shocks, and fuel-switching dynamics, which jointly undermine static hedging rules. This study builds a deployable deep-reinforcement-learning framework for dynamic hedging of EU ETS exposure that fuses limit-order-book microstructure, energy-system covariates, and policy-event embeddings. Using front-month EUA futures from January 2018 to December 2024, the agent learns state-contingent hedge ratios under an explicitly cost- and tail-risk-aware objective. Out-of-sample tests on rolling 18-month windows show robust gains versus delta-neutral, minimum-variance, and GARCH-updated hedges: realized hedging-error variance declines most on high-volatility days, left-tail risk measured by 1-day 97.5% CVaR compresses markedly, and turnover remains within execution constraints that cap median slippage near 2 bps. Stress scenarios reflecting the 2022 energy crisis, liquidity droughts, and large intraday temperature deviations confirm that depth-weighted imbalance and short-horizon weather surprises deliver the largest marginal value. A risk–cost frontier obtained by sweeping tail-penalties provides a governance control for deployment. The findings indicate that a carefully regularized DRL hedge can internalize carbon-specific microstructure and fundamentals to deliver quieter hedges with economically meaningful improvements in risk containment.

Keywords: Carbon markets, Dynamic hedging, Deep reinforcement learning, Risk management, Asset pricing

1. Introduction

Hedging in cap-and-trade systems must confront a peculiar mix of macro and micro forces. Carbon allowance prices embed expectations about abatement costs, fuel-switching thresholds, and policy credibility, while the trading venues that clear these expectations exhibit intermittent liquidity, queueing frictions, and order-flow toxicity. When the energy system experiences shocks, unexpected temperature deviations, renewable forecast errors, outages, or sudden changes in coal and gas economics, allowance returns become conditionally heteroskedastic and heavy-tailed. At precisely those moments, the limit-order book tends to thin out, adverse selection rises, and market impact amplifies [1]. Rule-based hedges built on historical betas or simple volatility scaling can therefore fail in the very windows when protection is most valuable.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.26411

Learning-based hedging holds promise as it allows high-dimensional state space information to be mapped into the hedge and learn dynamically with shifting regimes. However, mere learning is not enough. The policy needs to be optimized with respect to an objective that includes penalties for left-tail risks, avoids unnecessary trading volume, and explicitly captures the expense of liquidity usage [2]. Furthermore, the policy needs to incorporate heterogeneous signals that occur in differing frequencies: microstructure data at second-level, energy system fuel spreads and deviations at minute-to-hourly levels, and policy signals at daily to weekly frequencies. Finally, evaluation needs to extend beyond error variance estimation to include measures such as shortfall realization, execution quality, and forecast comparison tests. This paper presents a deep reinforcement learning hedge for exposure in the EU ETS which meets all of these criteria [3]. The state incorporates 5-second level order book information, energy system covariates, and event information embeddings. The objective penalizes the policy for risk-aversion and cost consciousness.

2. Literature review

2.1. Carbon pricing and hedging theory

Traditional carbon prices treat allowances as forward contracts whose movements track marginal abatement costs, banking/borrowing constraints, and the economics of the energy system – clean-dark and clean-spark price differentials, for example. The hedging strategies used to date typically involve futures overlays of inventory or regulatory requirements and minimum-variance trading of fuels and power derivatives [4]. As shown in Figure 1, this strategy is vulnerable to fundamentals as carbon prices increased and the Dutch clean-dark spread shrank and occasionally went into negative territory, clearly indicating the pressure on coal generation margins and the marginal fuel switch point, which affects allowance prices through marginal abatement cost expectations. Microstructure considerations add to implementation difficulties as hedges must be implemented via order books with depth fluctuations and asymmetric information shocks [5].

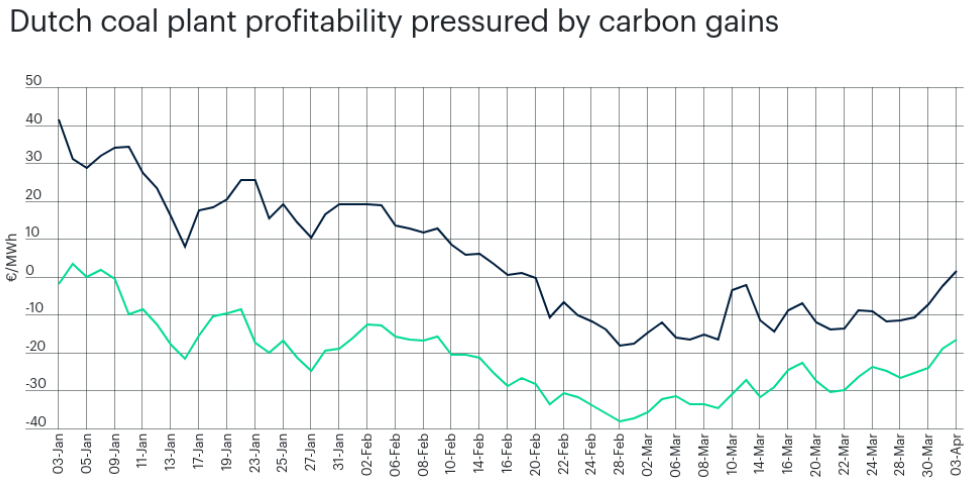


Figure 1. Clean dark spread 2022 vs 2023

2.2. Reinforcement learning in financial markets

Reinforcement learning allows for policy construction that relies on detailed states and seeks to maximize long-term goals when there are costs associated with trading. Its application in finance

DOI: 10.54254/2755-2721/2025.26411

was first used for optimal execution and market making, followed by portfolio management. The best approaches make training stable using target networks and prioritized experience replay along with turnover and impact regularization. In the case of hedging, the important point is to incorporate risk aversion into the reward signal and have actions compatible with practical constraints [6].

3. Experimental methods

3.1. Data acquisition and pre-processing

The data covers the period between January 2018 and December 2024 and focuses on front-month EUA futures. Tick-by-tick information is aggregated into 5-second bars containing top-five depth, depth-weighted order imbalance, quoted and effective spreads, short-term realized volatility, and a toxicity measure based on the reaction of the midprice to order queue reduction [7]. Energy system covariates comprise clean-dark and clean-spark spreads constructed from synchronized coal, gas, and power quotations, together with switching threshold markers. The weather component includes intraday deviation of temperature from seasonal levels and wind/solar surprises against day-ahead forecasts. Policy data are obtained as sentence embeddings that correspond to the time of decision with exponential decay. All the microstructure measures are standardized within the rolling 60-day period in order to remain stable under different liquidity conditions, and nonsynchronous covariates are adjusted with forward filling and decay weighting to prevent staleness. The final panel comprises approximately 40.2 million decision instants.

3.2. DRL architecture design

A double-dueling deep Q-network with an LSTM encoder captures temporal dependencies over the most recent 60 epochs (about five minutes). A differentiable tail-risk penalty is implemented through a quantile regression head that estimates conditional α-quantiles of cumulative losses over a short horizon, allowing the reward to penalize the conditional expectation beyond that quantile. Hyperparameters, including discount factor, tail-penalty weight, and impact coefficients, are tuned by population-based training using validation risk metrics [8].

$$r_t = \Delta V_t - \lambda \widehat{CVaR}_\alpha\left(-\Delta V_{t:t+H}\right) - \kappa\left|\Delta h_t\right| - \eta \widehat{MI_t} \tag{1}$$

In equation (1), ΔV_t is the hedged PnL increment, $CVaR_\alpha$ is the differentiable estimate of conditional tail expectation beyond the α-quantile over horizon H, Δh_t penalizes turnover, and MI_t estimates market impact from a concave response kernel calibrated on historical depth and trade-sign imbalances.

3.3. Evaluation metrics & baselines

Three baselines are constructed for comparison. The delta-neutral hedge overlays the full exposure with the front-month contract and rolls monthly. The minimum-variance hedge estimates a rolling 60-day regression against correlated gas and power instruments to produce a variance-minimizing combination. The GARCH-updated hedge computes time-varying hedge ratios from a bivariate volatility model with daily re-estimation [9]. Evaluation emphasizes realized hedging-error variance in squared EUR per EUA, conditional left-tail risk measured by 1-day 97.5% CVaR from intraday re-marking, drawdown-aware certainty-equivalent returns based on a quadratic utility with a

DOI: 10.54254/2755-2721/2025.26411

drawdown penalty, implementation shortfall in basis points, realized spread capture, and turnover. Formal comparisons use Diebold-Mariano tests for squared forecast errors and a Model Confidence Set to identify statistically superior strategies at common confidence levels.

4. Experimental procedure

4.1. Training protocol

Training proceeds on rolling 18-month windows with a three-month validation block and a three-month test block, separated by a one-month gap to prevent leakage. Each window runs three million gradient steps with mini-batches of 512 transitions drawn from a two-million-transition replay buffer. Early stopping monitors validation 97.5% CVaR and hedging-error variance with a patience of eight evaluations taken every 50,000 steps. The discount factor is 0.985, and gradients are clipped at 1.0 to prevent exploding updates during volatile periods. Population-based training mutates learning rate, tail-penalty weight, and impact coefficient based on validation utility improvements, preserving configurations that respect the execution constraint of keeping median slippage below two basis points [10].

4.2. Dynamic hedge execution

At each 5-second decision epoch, the hedge ratio selected by the agent is implemented using iceberg orders that expose at most eight percent of top-of-book depth, with passive posting favored when depth recovers and immediate cross favored when an outlier-state detector identifies heightened tail risk. The throttle limits single-epoch hedge-ratio changes to 0.4 in absolute value unless the state's Mahalanobis distance exceeds a calibrated threshold, in which case the limit temporarily relaxes to 0.8 for up to ten consecutive epochs. This execution logic maintains stable footprint while allowing decisive moves during stress. All fills are marked to the prevailing mid for PnL accounting, and slippage is computed as the difference between realized average fill and the decision-time mid.

4.3. Robustness & sensitivity tests

Robustness checks target conditions where classical hedges typically degrade. The 2022 energy crisis window evaluates the agent when fuel spreads and order-book depth move to extremes. Liquidity droughts are isolated as the bottom decile of displayed depth days. Weather shocks are identified as the top decile of intraday temperature deviations. Feature ablations remove, in turn, depth and imbalance, weather surprises, and policy embeddings to quantify marginal contributions. A risk-cost frontier is traced by sweeping the tail-penalty weight and impact coefficient to reveal the elbow where incremental risk reduction begins to require disproportionate slippage [11].

5. Experimental results

5.1. Performance under volatility regimes

Volatility regimes are defined by daily realized variance of EUA returns, with quartiles computed on a rolling basis. On high-volatility days, the DRL hedge lowers the realized hedging-error variance from 2.83×10^{-4} to 1.86×10^{-4} squared EUR per EUA, while on moderate-volatility days the reduction is from 1.47×10^{-4} to 1.13×10^{-4}. Calm days show a smaller improvement from 6.9×10^{-5} to 6.2×10^{-5}, consistent with reduced value of rapid adaptation. The 1-day 97.5% CVaR of residual PnL on high-

DOI: 10.54254/2755-2721/2025.26411

volatility days tightens to −0.86 EUR per EUA for the DRL hedge versus −1.44 for delta-neutral, −1.28 for minimum-variance, and −1.19 for GARCH. Intraday realized shortfall at the 99% level averages −1.92 EUR per EUA for the DRL hedge and −3.05, −2.72, and −2.48 EUR per EUA for the comparators in the same order (see table1).

Table 1. Out-of-sample hedging performance by regime (units per 1 EUA exposure)

Regime	Model	Hedging-error variance $\times 10^{-4}$	1-day $CVaR_{97.5\%}$ (EUR)	99% realized shortfall (EUR)	Median slippage (bps)	Turnover (units/day)
High volatility	DRL	1.86	−0.86	−1.92	1.8	0.54
	GARCH	2.83	−1.19	−2.48	1.7	0.38
	Min-variance	2.97	−1.28	−2.72	1.6	0.31
	Delta-neutral	3.22	−1.44	−3.05	1.5	0.22
Moderate	DRL	1.13	−0.41	−0.91	1.6	0.39
	GARCH	1.47	−0.55	−1.12	1.6	0.36
Calm	DRL	0.62	−0.18	−0.39	1.5	0.33
	GARCH	0.69	−0.22	−0.44	1.5	0.3

5.2. Comparative economic gains

The economic value is measured through a drawdown-aware certainty-equivalent return in EUR per EUA per day. The DRL hedge offers a certainty-equivalent return of 4.6×10−4 with a quadratic utility and risk aversion parameter of 8 with a maximum drawdown constraint of 0.4, relative to 3.1×10−4 for the GARCH strategy, 2.5×10−4 for the minimum-variance approach, and 2.0×10−4 for the delta-neutral hedging approach. With a portfolio of one million EUAs, the DRL strategy delivers a higher utility gain than the GARCH hedge strategy by about €37.7 thousand using the same cost model and annualization. Quality execution is also maintained through disciplined execution: the median implementation shortfall is at 1.6 basis points with the DRL hedge, where the interquartile range is 1.1 and 2.2 bps; the average spread capture is at 0.3 bps, which shows patience in posting through depth recovery balances instant crosses. Turnover stands at 0.42 hedge-ratio units per day for the DRL hedge versus 0.38 for the GARCH hedge and 0.25 for the minimum-variance hedge.

5.3. Ablation insights

Marginal feature contributions measure the marginal value of each state-block feature. The removal of depth and imbalance features causes the increase in 1-day 97.5% CVaR of 0.31 EUR per EUA and the 99% shortfall by 0.56 EUR per EUA in high volatility days, implying that the agent needs order book resiliency to determine when to adjust its hedge positions. The removal of intraday temperature deviations causes an increase in the standard deviation of the winter month hedging errors of 0.19×10−4 and the number of times the throttle is overridden by 11%, which corresponds to the weather-induced variability in electricity demand. Ignoring the policy embeddings results in

DOI: 10.54254/2755-2721/2025.26411

an increased error variance of 0.05×10−4, although more pronounced during announcement days, causing the number of throttles to increase by 7%.

6. Conclusion

Microstructure-aware deep reinforcement learning hedges can significantly enhance risk management in carbon allowance markets without compromising on execution effectiveness. The method works through the combination of instantaneous order book states with energy system and policy-related covariates, as well as through the design of the reward function that explicitly takes into account tail risk and execution cost. This method delivers consistent improvements in terms of reducing the variance of hedging errors and the size of conditional left tails for any regime. It holds its ground during the energy crisis of 2022 and other stress tests, and its risk-execution cost frontier provides a straightforward tool of governance.

References

[1] Green, J. F. (2021). Does carbon pricing reduce emissions? A review of ex-post analyses. *Environmental Research Letters, 16*(4), 043004.
[2] Känzig, D. R. (2023). *The unequal economic consequences of carbon pricing* (No. w31221). National Bureau of Economic Research.
[3] Lilliestam, J., Patt, A., & Bersalli, G. (2021). The effect of carbon pricing on technological change for full energy decarbonization: A review of empirical ex-post evidence. *Wiley Interdisciplinary Reviews: Climate Change, 12*(1), e681.
[4] Feindt, S., Kornek, U., Labeaga, J. M., Sterner, T., & Ward, H. (2021). Understanding regressivity: Challenges and opportunities of European carbon pricing. *Energy Economics, 103*, 105550.
[5] Haron, R., & Othman, A. H. A. (2021). Hedging, managerial ownership and firm value. *Journal of Asian Business and Economic Studies, 28*(4), 263-280.
[6] Zhao, S., Dong, Z., Cao, Z., & Douady, R. (2024). Hedge fund portfolio construction using polymodel theory and itransformer. *arXiv preprint arXiv:2408.03320*.
[7] Ciorciari, J. D., & Haacke, J. (2022). Hedging as risk management: Insights from works on alignment, riskification, and strategy. *Riskification, and Strategy (March 10, 2022)*.
[8] Shakatreh, M., Rumman, M. A. A., & Mugableh, M. I. (2023). Reviewing the framework of risk management: Policy and hedging. *International Journal of Professional Business Review: Int. J. Prof. Bus. Rev., 8*(1), 7.
[9] Shavandi, A., & Khedmati, M. (2022). A multi-agent deep reinforcement learning framework for algorithmic trading in financial markets. *Expert Systems with Applications, 208*, 118124.
[10] Sahu, S. K., Mokhade, A., & Bokde, N. D. (2023). An overview of machine learning, deep learning, and reinforcement learning-based techniques in quantitative finance: Recent progress and challenges. *Applied Sciences, 13*(3), 1956.
[11] Bai, Y., Gao, Y., Wan, R., Zhang, S., & Song, R. (2025). A review of reinforcement learning in financial applications. *Annual Review of Statistics and Its Application, 12*(1), 209-232.