Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

The illusion of predictability in carbon price forecasting and look-ahead bias

Ozan Nadirgil

onadirgil@proton.me

Independent https://orcid.org/0009-0003-5584-0488

Research Article

Keywords: Carbon pricing; EU ETS; Forecast evaluation; Machine learning; Information leakage; Fuel switching

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11861253/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

The illusion of predictability in carbon price forecasting and look-ahead bias

Ozan Nadirgil

Highlights

- No model beats a random walk for EUA returns at 1, 5 or 22 days in 2022 to 2026.
- Publication lags, purged refits and frozen tuning rule out leakage of future data.
- Drift learned in the 2018 to 2021 rally drives most long-horizon losses.
- Past EUA returns and policy uncertainty add small but consistent information.
- Same-day predictors would create spurious gains of up to 10% of squared error.

Abstract

Machine learning studies report large gains in forecasting EU Allowance (EUA) prices, but their designs rarely rule out leakage of future data. EUA futures returns are forecast one, five and twenty-two trading days ahead over 2022 to 2026 with nine models: a random walk, an autoregressive model with exogenous inputs, an elastic net, Extra Trees, XGBoost, k-nearest neighbours, support vector regression, a feed-forward network and an LSTM. Nineteen daily series enter in four economically motivated blocks: energy fundamentals, including the coal-to-gas switching price, financial markets, bond yields, and uncertainty and market activity. Every predictor is observable at the forecast origin, hyperparameters are fixed on 2018 to 2021 data, and models are refitted monthly with a purge between training targets and the origin. No model beats a random walk at any horizon; Clark and West tests, Holm-adjusted Diebold and Mariano tests and the model confidence set agree. Most long-horizon loss reflects drift learned during the 2018 to 2021 rally. Past EUA returns and policy uncertainty carry small but consistent OOS information, whereas the renewable share and the switching gap, prominent in in-sample rankings, impair forecasts. Measuring predictors on the target day would have produced spurious gains of up to ten percent.

Keywords: Carbon pricing; EU ETS; Forecast evaluation; Machine learning; Information leakage; Fuel switching

JEL classification: C53; C58; G17; Q54

1. Introduction

Roughly 40 percent of the EU's greenhouse gas emissions are priced through the EU Emissions Trading System (EU ETS); its allowance, the EUA, is now among the most closely followed commodities in Europe. Utilities hedge it alongside fuel and steelmakers and cement producers carry it on their balance sheets. Since 2018 the price has also attracted investment funds, which now hold sizeable net positions in EUA futures (1). A price that matters to this many agents invites forecasting, and the empirical literature has responded with a steady flow of models borrowed from machine learning.

Methodological scrutiny, however, has not kept pace. Much of the recent work reports gains over simple benchmarks that are large by any standard of financial forecasting, yet the forecasting set-up behind those gains is often hard to reconstruct. Predictors are sometimes dated to the very day of the target. Feature selection is frequently carried out on the full sample, evaluation window included. Model rankings rest on a single train and test split, and the statistical comparison of many models at once is seldom corrected for multiple testing. These issues are not cosmetic. Each of them biases reported accuracy upward, and together they can turn an unpredictable series into an apparently predictable one (2, 3).

A second gap is conceptual. The allowance is not an ordinary asset. Where permits can be banked, a capped market should price them at the marginal cost of meeting the cap over time, which links the price to fuel prices through coal-to-gas switching, to interest rates through the cost of carrying banked allowances, and to expected output through the demand for emissions (4, 5, 6). A forecasting study that ignores these channels can rank models, but it cannot say much about why one set of predictors helps and another does not. That limits its usefulness to those who act on the results.

This paper addresses both gaps. EUA futures returns are forecast at horizons of one, five and twenty-two trading days over January 2018 to October 2026, using nineteen daily series grouped into four economically motivated blocks, including energy fundamentals, financial markets, bond yields, and uncertainty and market activity. The sample begins with the 2018 reform that made the Market Stability Reserve (MSR) operational, so it covers the late years of Phase III, the whole of Phase IV to date, and the 2022 energy crisis. Every predictor dated t is observable at the close of day t, after publication lags. Hyperparameters are tuned once, on 2018 to 2021 data, and then frozen. Models are refitted every month on an expanding window, and the training sample at each refit excludes observations whose targets were not yet known at the forecast origin. Accuracy is judged against a random walk (RW), an autoregressive model with exogenous regressors and an

elastic net, using the out-of-sample (OOS) R^2, the Clark and West test, Diebold and Mariano tests with Holm correction and the model confidence set of Hansen, Lunde and Nason (7, 8, 9, 10).

The model set is deliberately compact. Instead of stacking variants of the same algorithm, the study retains one representative per family, so that each comparison answers a distinct question. Extra Trees and XGBoost represent bagging and boosting of regression trees. k-nearest neighbours asks whether days that resemble each other in predictor space are followed by similar returns, and support vector regression adds a kernel method with a long record in forecasting carbon prices. A feed-forward network and a recurrent LSTM network (long short-term memory) share the same inputs and differ only in their recurrent memory, which makes it possible to test whether past sequences add anything beyond the current state.

The paper makes three contributions. First, it provides an OOS benchmark for EUA forecasting that is free of look-ahead bias by construction and that uses tests suited to many overlapping comparisons. The aim is not to identify a single best model but to establish what level of predictability survives a strict protocol, a step the field arguably needs before more complex hybrids are proposed. Second, it ties predictor selection to the theory of allowance pricing. The coal-to-gas switching price and the gap between the EUA and that price are constructed from fuel and exchange rate data, a daily measure of German wind and solar output is added, and predictive relevance is measured by block as well as by variable, in the test sample and with a common metric across models. Third, it documents how predictive relevance shifts between the crisis years 2022 to 2024 and the calmer period that followed, which speaks directly to the debate on whether allowance prices are driven by fundamentals or by financial trading (1, 11).

The main result is negative, yet informative. Over January 2022 to October 2026, no model forecasts EUA returns better than the RW at any horizon, and none of the differences in favour of a learner is statistically significant once nested models and multiple comparisons are taken into account. Much of the loss at longer horizons comes from the strong drift that every model learns from the 2018 to 2021 rally. Within this overall picture, the EUA's own recent returns and policy uncertainty carry small but consistent information at weekly and monthly horizons, fuel price returns help at the monthly horizon, and structural state variables, including the renewable share and the switching gap, although prominent in in-sample importance rankings, damage OOS forecasts. A simple experiment shows that dating predictors to the day of the target would have turned these negative results into apparent gains of up to ten percent of the squared error.

The term "predictive relevance" is used deliberately. Permutation measures show how much a forecast deteriorates when a predictor is scrambled. They do not identify causal effects, and causal language is avoided throughout.

The paper is structured as follows. The economic channels behind the predictor set come first (Section 2), followed by the related empirical work (Section 3), the data (Section 4) and the forecasting design (Section 5). Sections 6 to 8 present, interpret and summarise the findings.

2. Allowance price formation and the choice of predictors

2.1. Intertemporal arbitrage and the cost of carry

EU allowances can be banked without limit from one trading phase to the next. Under banking and perfect foresight, the expected allowance price rises at the rate of interest, as in the Hotelling rule for an exhaustible resource (4). A futures contract on an allowance is then priced off the spot price plus the cost of financing the position until delivery. Two implications follow for forecasting. Short-term interest rates enter the relation between spot and futures prices directly. Long-term rates matter through the discount rate that firms apply when they decide how many allowances to hold for future compliance. Empirical work on the European market finds that firms behave as if they use much higher discount rates and shorter planning horizons than a pure Hotelling model would imply (12), so the link between rates and allowance prices may be weaker and less stable than theory suggests. The German three-month and ten-year yields are therefore included, and the strength of the rate channel is treated as an empirical question.

2.2. Fuel switching and the marginal abatement cost

In the short run, the cheapest way for the power sector to cut emissions is to run gas plants instead of coal plants. The allowance price at which a gas unit and a coal unit have equal marginal costs is the fuel-switching price (13):

$$FSP_t = \frac{\frac{P_{gas,t}}{\eta_{gas}} - \frac{P_{coal,t}}{\eta_{coal}}}{\frac{EF_{coal}}{\eta_{coal}} - \frac{EF_{gas}}{\eta_{gas}}} \quad [1]$$

where P_{gas} and P_{coal} are fuel prices per MWh of thermal input, η denotes plant efficiency and EF the emission factor per MWh of fuel. When the EUA trades above the switching price, gas plants are cheaper than coal plants and switching supplies abatement. When it trades below, coal stays in merit and emissions remain higher. The switching price is therefore a natural anchor for the short-run marginal abatement cost, and its rise in 2021 and 2022, driven by gas, explains part of the decoupling between gas and carbon in those years (Fig. 1b).

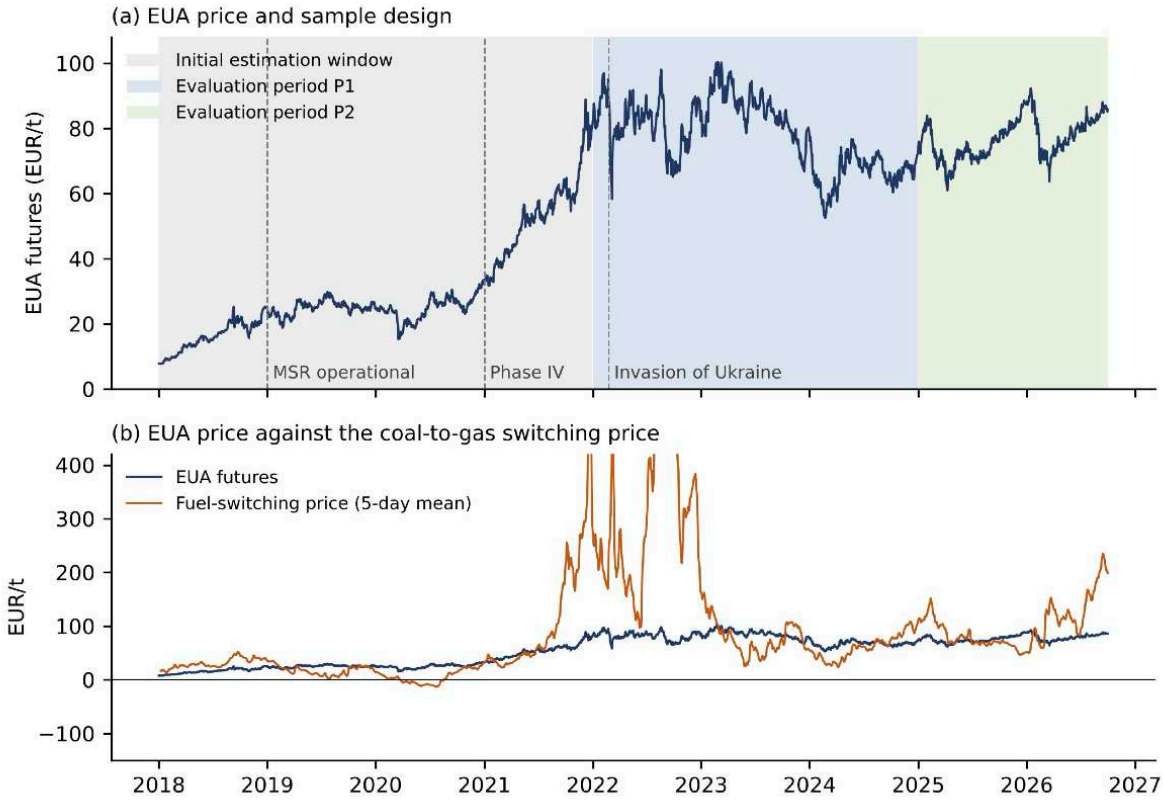


Fig. 1. Price of the December EUA futures contract, 2018 to 2026. Panel (a) marks the initial estimation window and the two evaluation sub-periods; panel (b) plots it against the coal-to-gas fuel-switching price of Eq. [1] (five-day mean; values above €420 in 2022 are truncated).

The switching price is computed daily from TTF gas, API2 coal converted into euros, and standard plant parameters (η_{gas} = 0.50, η_{coal} = 0.38, EF_{gas} = 0.202 and EF_{coal} = 0.341 tCO2 per MWh thermal). The switching gap, the difference between the EUA and the switching price, enters as a state variable. A large positive gap means that allowances are expensive compared with the abatement that fuel switching can deliver, which may signal pressure for a correction. Electricity prices and the share of wind and solar in German load complete the block. Higher renewable output pushes fossil plants down the merit order and lowers expected emissions, which should weigh on allowance demand (14).

2.3. Economic activity, risk appetite and the financialisation of the market

Allowance demand follows industrial output and power generation, so expectations about activity should feed into allowance prices. Equity indices summarise those expectations at a daily frequency. The predictor set covers the broad European market (STOXX Europe 600), European utilities, which are the largest compliance buyers, the main US large-cap index as a gauge of global conditions, and an index of green energy stocks that captures expectations about the pace of

decarbonisation. The slope of the US Treasury curve is included as a leading indicator of recessions (15).

Since the 2018 reform the EUA has also become a financial asset. Investment funds take directional positions, and the price now reacts more strongly to risk sentiment and to liquidity conditions (1, 6). Two consequences follow for forecasting. Gauges of risk appetite like the VIX may carry information about allowance returns that has little to do with abatement costs. And trading activity in the EUA market itself, measured by volume and intraday range, may help predict short-term moves through order flow and volatility dynamics. The euro to dollar exchange rate enters because coal and oil are priced in dollars while allowances are priced in euros.

2.4. Policy and geopolitical uncertainty

The ETS is a market created by policy. The MSR, the Fit for 55 reforms and the REPowerEU plan all shifted expected supply, and in 2022 the war in Ukraine changed the fuel mix of European power generation within months. News-based indices offer a daily view of this kind of uncertainty. This study employs the geopolitical risk index of Caldara and Iacoviello (16) and the US economic policy uncertainty index of Baker, Bloom and Davis (17). Neither is specific to carbon. The question of interest is whether broad uncertainty shocks, which affect fuel markets and asset markets alike, also carry predictive content for allowance returns.

2.5. What the channels imply for forecasting

These channels do not translate into precise point predictions. They suggest, however, a pattern that the data can confirm or reject. If the allowance price is anchored by fundamentals, energy variables and the switching gap should carry the bulk of the predictive content, and their relevance should rise with the forecast horizon as short-term noise averages out. If the market behaves mainly as a financial asset, equity, volatility and activity measures should dominate, especially at short horizons. And if the market is close to weak-form efficient, no block should deliver forecasts that beat a random walk by a margin that survives multiple testing (18). Section 6 evaluates these possibilities in turn.

3. Related literature

3.1. Drivers of allowance prices

Early empirical work on the European carbon market asked which fundamentals move the allowance price. For the first trading phase, Hintermann (5) finds that prices were unrelated to marginal abatement costs until the first compliance data showed the cap to be slack; afterwards,

fuel prices and weather explained more of their movement. Aatola et al. (19) estimate a model of market fundamentals and report a strong role for electricity, gas and coal prices. Koch et al. (14) study the long price decline after 2008 and find that economic activity, renewable deployment and offset credits explain only a small part of it, which leaves room for expectations and policy credibility. Structural models of the gas, power and carbon nexus agree: gas prices move allowance prices, and both are passed through to electricity (20).

Later studies widened the list of candidate drivers as well as the toolkit. Multiscale error-correction work singles out power, coal and equity returns, each at its own frequency (21). Quantile regressions find that fuel prices matter most in the tails of the distribution of carbon prices (22). During the pandemic, fossil fuels and utility shares mattered most, and COVID-19 reshaped how these links worked (23). Connectedness studies trace time-varying spillovers into carbon from fuels, equities and other commodities (24), and a wavelet study reports co-movement of carbon with green bonds that differs by time scale (25). Jumps, outliers and oil market risk also help to explain carbon volatility (26).

3.2. Policy, banking and financial trading

The 2018 reform changed the analytical picture. The MSR withdraws allowances when the surplus is large and, since 2023, invalidates part of the reserve. Theory predicts that such a mechanism shifts scarcity forward in time and makes prices react more to expectations about future policy (27, 28). Friedrich et al. (11) test whether the price surge of 2018 reflected fundamentals or speculative dynamics and find signs of explosive behaviour that fundamentals alone do not explain. A review by the same group organises the literature around demand fundamentals, regulatory intervention and financial trading, and argues that the third strand has become central since the reform (6). Quemin and Pahle (1) show that investment funds have expanded their positions and warn that speculation can affect price formation. Evidence from other asset classes points the same way: shocks travel between green equities, technology stocks, commodities and currencies, and the channels shift during crises (29).

3.3. Forecasting with machine learning

Forecasting studies have moved from GARCH-type volatility models to machine learning and hybrid decomposition methods. News-based attention to climate change improves predictions of carbon futures returns (30). A comparison of factor and combination methods for EUA volatility finds bond and commodity predictors informative (31). Adding technology indices to a wide pool of learners, another study reports that boosted trees do best at all horizons (32). A less common

conclusion comes from work that selects predictors with care (33): there, linear models beat non-linear ones in small samples because the latter overfit. Decomposition hybrids, which split the price series into components before modelling each one, dominate many recent comparisons (34, 35). For Chinese pilot markets, connectedness with energy-intensive sectors appears weak and episodic (36).

3.4. What a credible forecasting comparison requires

Two lessons from the broader forecasting literature are directly relevant. The first concerns the benchmark. In exchange rate and equity forecasting, the RW and the historical mean have proved hard to beat in genuine forecasting exercises, and in-sample fit is a poor guide to forecasting value (37, 38, 39). Gains over a random walk should therefore be small if markets are reasonably efficient, and large reported gains call for a close look at the design. The second concerns the evaluation itself. Standard cross-validation leaks information when observations are serially dependent, and choosing predictors with data that later form the evaluation sample inflates measured accuracy (2). Machine learning studies in empirical finance address this with expanding windows, strict separation of tuning and test samples and tests that account for nested models (3, 7).

Set against these lessons, the carbon forecasting literature has three gaps. Many studies use contemporaneous predictors or full-sample feature selection, which makes their accuracy figures hard to compare with a genuine forecast. Few report multiple-testing corrections when many models are compared at once. And few connect the predictors they find important to the economic channels of Section 2. The design adopted here addresses all three.

4. Data

4.1. Sample and sources

The sample runs from 2 January 2018 to 1 October 2026, which gives 2,251 trading days of the EUA futures market. It starts in the year in which the EU adopted the reform directive that made the MSR operational from January 2019 (40). The end date is the last day on which all market series were available when the data were collected. Table 1 lists the variables, their sources and the transformations used in estimation.

Table 1. Variables, transformations and sources.

Code	Variable	Block	Transformation	Source
EUA	EUA futures, December contract (EUR/t)	Own dynamics	Δlog (1, 5, 22 days)	Investing.com
TTF	Dutch TTF natural gas, front month (EUR/MWh)	Energy	Δlog (1, 5 days)	Investing.com
COAL	API2 Rotterdam coal, front month (USD/t, in EUR)	Energy	Δlog (1, 5 days)	Investing.com
BRENT	Brent crude oil, front month (USD/bbl)	Energy	Δlog (1, 5 days)	Investing.com
ELC	German day-ahead electricity price (EUR/MWh)	Energy	Δ (1, 5 days)	Ember (ENTSO-E)
RES	Wind and solar output, share of German load (%)	Energy	5-day mean; Δ (1 day)	Energy-Charts
FSP	Coal-to-gas switching price (EUR/t), Eq. [1]	Energy	Δ (1, 5 days)	Derived
GAP	EUA minus switching price (EUR/t)	Energy	Level	Derived
STOXX	STOXX Europe 600	Financial	Δlog (1, 5 days)	Investing.com
UTIL	STOXX Europe 600 Utilities	Financial	Δlog (1, 5 days)	Investing.com
SPX	S&P 500	Financial	Δlog (1, 5 days)	Investing.com
NCE	NASDAQ Clean Edge Green Energy	Financial	Δlog (1, 5 days)	Investing.com
VIX	CBOE Volatility Index	Financial	Log level; Δlog (5 days)	Investing.com
EURUSD	Euro to US dollar exchange rate	Financial	Δlog (1, 5 days)	Investing.com
DE10Y	German 10-year government bond yield (%)	Rates	Δ in bp (1, 5 days)	Investing.com
DE3M	German 3-month government bill yield (%)	Rates	Δ in bp (1, 5 days)	Investing.com
USSPR	Treasury spread, 10-year less 2-year, United States (pp)	Rates	Δ in bp (1, 5 days)	FRED (T10Y2Y)
GPR	Daily geopolitical risk index	Uncertainty	Log, 5-day mean, lag 5	Caldara and Iacoviello
EPU	US daily economic policy uncertainty index	Uncertainty	Log, 5-day mean, lag 1	FRED (USEPUINDXD)
VOL	EUA futures trading volume	Uncertainty	Log ratio to 20-day mean	Investing.com
RANGE	EUA intraday range (Parkinson estimator)	Uncertainty	Level; 5-day mean	Investing.com

Notes: Δlog denotes log returns in percent and Δ level differences. Energy: energy fundamentals; Financial: financial markets; Rates: interest rates; Uncertainty: uncertainty and market activity. Lags refer to trading days and reflect publication schedules. Clean dark and clean spark spreads were considered and dropped because their changes correlate above 0.95 with the electricity price.

The target is the continuous series of the December (yearly) EUA futures contract, which is the most liquid point of the curve and the reference for hedging. Fuel prices are front-month futures for TTF gas, API2 Rotterdam coal and Brent crude oil. Coal is converted from dollars per tonne to euros per MWh of thermal input using the EUR/USD rate and an energy content of 6.978 MWh per tonne. German day-ahead power prices are Ember's load-weighted daily averages of ENTSO-E data. The renewable share is the sum of onshore wind, offshore wind and solar generation as a share of German load, taken from the Energy-Charts platform of Fraunhofer ISE. Equity indices, VIX, EUR/USD and German government bond yields come from Investing.com, and the spread between US Treasury yields at 10 and 2 years and the daily economic policy uncertainty index come from FRED. The geopolitical risk index is the daily series of Caldara and Iacoviello (16).

4.2. Alignment, cleaning and information timing

All series are aligned to EUA trading days by calendar date. When a market is closed on an EUA trading day, its last observation is carried forward for at most three days; this affects fewer than 80 observations for any series and at most four for the energy variables. Two isolated prints in the

German three-month yield (6 and 11 April 2023) lie far outside the surrounding values and are treated as missing. Continuous front-month futures series contain jumps on contract roll dates. Rather than modelling the rolls, every predictor is winsorised at the 0.5th and 99.5th percentiles, with cut-offs estimated on the training sample at each refit, so that no information from the evaluation window enters the transformation.

Timing deserves particular care. Every predictor dated t is known at the close of trading on day t. The two news-based indices are lagged to reflect their publication schedules: the policy uncertainty index by one day and the geopolitical risk index, which is updated weekly, by five trading days. The EUA enters through its own past returns, a one-month momentum term and the intraday range.

4.3. Transformations and properties

Prices that are always positive enter as one-day and five-day log returns in percent. Electricity prices, which turn negative on some days, and the switching price enter as level differences. Bond yields and the US spread enter as changes in basis points. The VIX enters in logs, the renewable share as a five-day average and as a daily change, and the switching gap in levels. Abnormal volume is the log ratio of today's volume to its average over the previous 20 days, and the intraday range is the Parkinson estimator (41). The final design matrix has 39 columns built from 22 underlying variables.

Table 2. Descriptive statistics and unit root tests for the transformed predictors.

Variable	Form	Mean	SD	Skewness	Ex. kurtosis	ADF (level)	ADF (transformed)
EUA	Δ	0.100	2.626	−0.50	5.23	−1.53	−21.01***
TTF	Δ	0.064	5.093	0.44	8.01	−2.34	−21.08***
COAL	Δ	0.026	2.995	−2.20	69.33	−1.83	−20.21***
BRENT	Δ	0.018	2.644	−1.24	15.78	−2.21	−19.86***
ELC	Δ	0.074	29.364	−0.42	9.24	−3.62***	−13.76***
RES	Δ	−0.003	13.196	0.07	0.83	−4.32***	−15.49***
FSP	Δ	0.083	18.002	−1.79	54.08	−2.93**	−9.48***
GAP	level	−37.213	108.118	−3.47	15.22	−3.30**	−3.30**
STOXX	Δ	0.021	0.988	−1.28	16.39	−0.10	−10.57***
UTIL	Δ	0.028	1.096	−1.64	20.33	−1.15	−11.44***
SPX	Δ	0.045	1.194	−0.66	15.55	0.62	−13.11***
NCE	Δ	0.038	2.371	−0.21	3.08	−1.96	−11.17***
VIX	level	2.929	0.298	0.98	1.57	−5.48***	−5.93***
EURUSD	Δ	−0.004	0.445	0.11	1.72	−2.49	−17.62***
DE10Y	Δ	0.127	4.673	−0.23	4.10	0.22	−9.52***
DE3M	Δ	0.141	3.733	1.19	14.96	−0.63	−8.59***
USSPR	Δ	−0.005	3.762	0.73	8.46	−1.53	−9.41***
GPR	level	4.783	0.384	0.37	0.31	−5.30***	−3.62***
EPU	level	4.967	0.599	0.43	−0.59	−3.05**	−2.54
VOL	abn.	−0.044	0.366	−1.14	5.34	−5.70***	−11.42***
RANGE	level	2.185	1.203	2.88	18.43	−6.65***	−6.65***

Notes: Statistics cover the estimation series used in the models, 1 February 2018 to 1 October 2026 (2,229 observations). Δ: one-day transformation as in Table 1; level: level form used in the models; abn.: log volume relative to its 20-day mean. ADF statistics (with a constant and an AIC-selected lag length) are computed for the level series (prices, yields, indices) and for the transformed series. ***, ** and * mark significance at 1, 5 and 10 percent.

Table 2 reports summary statistics for the transformed series. Every transformed predictor rejects a unit root at 1 percent, with two exceptions: the policy uncertainty index, whose ADF statistic is borderline, and the switching gap, which is stationary at 5 percent. Returns are fat-tailed, most strongly for coal, where roll dates produce large one-day moves. Correlations among the 39 predictors are modest: 92.7 percent of all pairs have an absolute correlation below 0.3, the largest is 0.77 (five-day European and US equity returns), and no variance inflation factors (VIFs) exceed 5.9. An earlier specification with clean dark and clean spark spreads produced correlations above 0.95 with the electricity price, because both spreads are dominated by power price movements; they were dropped for that reason.

5. Forecasting design

5.1. Target and information set

Let P_t be the EUA futures settlement price on day t. The target at horizon h is the cumulative log return over the next h trading days,

$$y_{t,h} = 100\,(ln\ P_{t+h} - ln\ P_t) \qquad [2]$$

for $h \in \{1, 5, 22\}$. Forecasts are made at the close of day t using the predictor vector x_t described in Section 4. Returns rather than price levels are forecast, for two reasons. Tree-based learners cannot extrapolate beyond the range of the training targets, which makes level forecasts meaningless when the price moves into a new range, as it did in 2021. And a return forecast of zero is exactly the RW forecast for the price, which gives a transparent benchmark. Price forecasts are recovered as $\hat{P}_{t+h} = P_t \exp(\hat{y}_{t,h}/100)$ and are used for the euro-denominated error measures.

5.2. Models

The benchmark set has three members. A random walk (RW) forecasts a zero return. The ARX model regresses the target by ordinary least squares on recent EUA returns and on the one-day changes and level terms of the other predictors. The elastic net (ENET) uses the full predictor set with a combined L1 and L2 penalty, which handles the correlated predictors by shrinking their coefficients together (42).

The machine learning set has one member per family. The Extra Trees model (ET) averages many regression trees grown on random split points and random subsets of predictors (43). The extra randomisation reduces variance relative to a random forest (44) and makes the ensemble less

sensitive to correlated inputs, because no single predictor dominates the initial splits. XGBoost (XGB) adds shallow trees one at a time, each fitted to the residual errors of the current ensemble, with shrinkage and subsampling (45); it is often the strongest learner on tabular data, but it can overfit noisy targets when trees are deep. k-nearest neighbours (KNN) forecasts the average target of the k most similar past days in standardised predictor space (46). It imposes no functional form, which is an advantage if the relation between predictors and returns differs across regimes, and a weakness when many irrelevant predictors dilute the distance metric. The feed-forward network (DNN) has two hidden layers (47). The LSTM network (48) receives the same predictors as a sequence covering the last ten trading days up to day t. Because the two networks differ only in their use of past sequences, comparing them isolates the value of recurrent memory. Support vector regression (SVR) with a radial basis kernel (49) represents kernel methods. It fits a function that ignores errors smaller than a tolerance band, which gives it some robustness to small noisy returns, but its accuracy depends heavily on the kernel width and the penalty parameter.

The families differ in how they react to correlated predictors, which is why their sensitivity is reported explicitly. Linear models suffer from unstable coefficients, which the elastic net penalty is designed to contain. Tree ensembles are largely unaffected in accuracy but split importance between correlated predictors, which biases split-based importance measures (50). KNN, SVR and the networks react to the scale of inputs, so all inputs are standardised with training-sample moments. Since all VIFs in the design are below 6, collinearity is a secondary concern for accuracy, and its effect on importance is addressed through the grouped permutation scheme of Section 5.5.

5.3. Estimation protocol

The protocol fixes every choice before the evaluation period begins (Fig. 2).

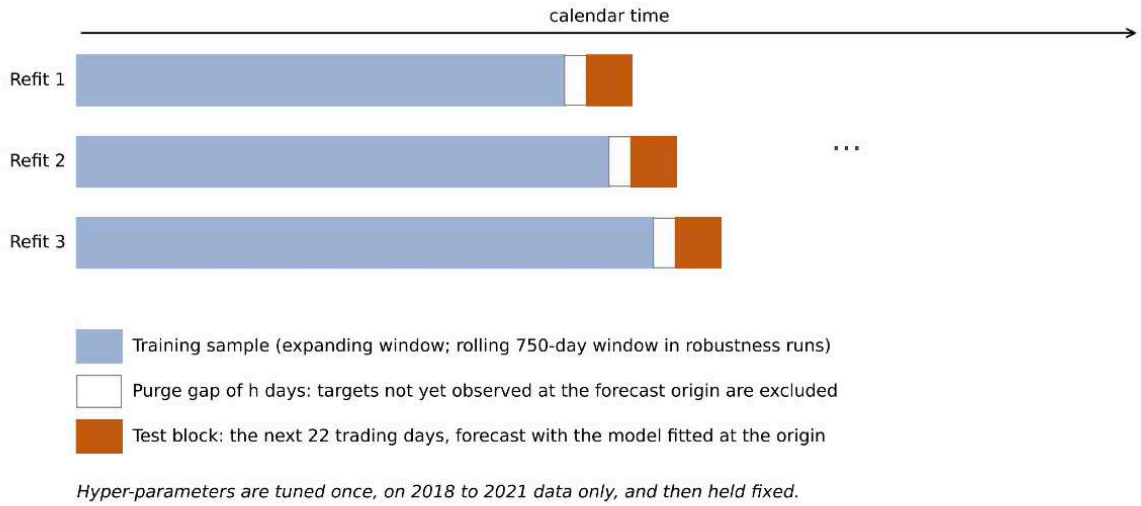


Fig. 2. Walk-forward forecasting design. Models are refitted every 22 trading days on all data available at the origin, excluding the last h days whose targets are not yet observed.

Hyperparameters are selected once, using data from February 2018 to December 2021 only. The scikit-learn models are tuned on four expanding folds in time order, separated by a gap of h days, so that no validation target overlaps a training target. For the two networks, the last 20 percent of this window is held out, again separated by h days. The search grids are small on purpose and are listed in Appendix A together with the selected values. The networks use the Adam optimiser (51), early stopping on the last 15 percent of each training sample, and an average of three random seeds.

The evaluation period starts on 3 January 2022. Every 22 trading days, each model is refitted on all observations from February 2018 up to the forecast origin and then produces forecasts for the next 22 days. When the forecast origin is day T, the training sample ends at $T - h$, because the targets of later days involve prices that are not yet observed at T. Omitting this purge would leak up to $h - 1$ days of future returns into training. Winsorisation bounds and standardisation moments are recomputed at each refit from the training sample alone. A rolling window of 750 days replaces the expanding window in a robustness check.

Results are reported for the full evaluation period and for two sub-periods: P1 (2022 to 2024), which contains the energy crisis and its unwinding, and P2 (January 2025 to October 2026). No modelling decision was taken with reference to either sub-period.

5.4. Forecast evaluation

Point accuracy is measured with the RMSE and the MAE (root mean squared and mean absolute errors) of return forecasts, with the RMSE of the implied price forecasts in euros, and with the OOS R^2 against the RW (39),

$$R^2_{OS} = 1 - \frac{\Sigma_t \left(y_{t,h} - \hat{y}_{t,h}\right)^2}{\Sigma_t y^2_{t,h}} \quad [3]$$

which is positive when a model beats the RW and is the natural metric when the benchmark forecast is zero. Theil's inequality coefficient U2, computed on prices, equals the ratio of a model's price RMSE to that of the RW. Directional accuracy is the share of correctly predicted signs, and its significance is tested with the Pesaran and Timmermann test (52).

Because every model nests the zero forecast, a standard Diebold and Mariano comparison is biased against the larger model when it is compared with the RW. The Clark and West test (7) is therefore applied; it adjusts the loss differential for the noise introduced by estimating parameters that are zero under the null:

$$f_t = y^2_{t,h} - \left[\left(y_{t,h} - \hat{y}_{t,h}\right)^2 - \hat{y}^2_{t,h}\right] \quad [4]$$

The statistic is the t-ratio of the mean of f_t, with a Newey and West variance using $h - 1$ lags to account for the overlap of multi-day targets, and is compared with one-sided normal critical values.

Pairwise comparisons among all models rely on the Diebold and Mariano test with the small-sample correction of Harvey, Leybourne and Newbold (8, 53):

$$DM^* = \frac{\bar{d}}{\sqrt{\frac{\hat{V}}{n}}} \sqrt{\frac{n + 1 - 2h + h(h-1)/n}{n}} \quad [5]$$

where $\bar{d}$ is the mean squared-error differential, $\hat{V}$ its Newey and West long-run variance and n the forecast count. Each horizon involves 36 pairwise tests, so p-values are adjusted with the Holm step-down procedure (9). Finally, the model confidence set (MCS) of Hansen et al. (10) identifies the subset of models that cannot be distinguished from the best at the 10 percent level. It uses a block bootstrap with 2,000 replications and a block length of max(h, 5) days, and it handles the multiple comparison problem by construction.

5.5. Predictive relevance of variables and blocks

The predictive relevance of a variable is measured by the rise in OOS mean squared error when that variable is scrambled. For every model and refit, the rows of all columns that belong to variable j are permuted across the evaluation period, the other predictors are kept fixed, and the forecasts are recomputed with the model fitted at that refit:

$$VI_j = 100 \times \frac{1}{R} \sum_{r=1}^{R} \frac{\Sigma_t \left(e_t^{(j,r)}\right)^2 - \Sigma_t e_t^2}{\Sigma_t e_t^2} \quad [6]$$

where e_t is the original forecast error, $e_t^{(j,r)}$ the error after the r-th permutation, and R = 5. Several aspects of this design matter. The measure is computed on test data only, so it reflects forecasting value rather than in-sample fit. It uses the same loss for every model, which makes the scores comparable across families, unlike split-based or weight-based importances, whose scales differ by orders of magnitude. And because all transformations of a variable (for example its one-day and five-day returns) are permuted together, correlated transformations of one variable cannot mask each other. The same procedure is applied to whole blocks, which also absorbs correlation between variables of the same block.

Variable rankings are aggregated across the eight learning models with a Borda count, where the best of K variables earns K points and the last one point. Because Borda scoring discards magnitudes, the mean relevance score and the mean rank are also reported, and the stability of rankings is measured across models and across sub-periods with Kendall's τ. Finally, to identify the direction of the effects learned by the tree ensembles, SHAP values (54) are computed for ET and XGB fitted on data up to December 2024, together with the correlation between each predictor and its SHAP contribution.

6. Results

6.1. Forecast accuracy against the RW

Table 3 sets out the accuracy of every model over the full evaluation period, January 2022 to October 2026. No model produces a positive OOS R^2 at any horizon. At the daily horizon the losses are small: the elastic net comes closest to the RW with an R^2_{OS} of −0.47 percent, the machine learning models lie between −1.29 percent (SVR) and −3.64 percent (XGB), and the ARX model trails at −5.81 percent. In price terms, the RW misses the next-day settlement by €1.78 (RMSE); the elastic net adds less than a cent to that error and the best machine learning model about one cent. Losses widen with the horizon. At five days the R^2_{OS} ranges from −2.09 percent (ET) to −10.79 percent (XGB), and at twenty-two days from −12.33 percent (ET) to −33.10 percent (ARX).

Table 3. OOS forecast accuracy, January 2022 to October 2026.

Model	R^2_{OS} (%)	RMSE	MAE	U2	DA (%)	CW p	DM p (Holm)	MCS p
Panel A: h = 1								
RW	0.00	2.318	1.683	1.000	.	.	.	1.000
ARX	−5.81	2.384	1.736	1.030	50.1	0.459	0.056	0.036†
ENET	−0.47	2.323	1.689	1.003	49.6	0.599	1.000	0.275
ET	−2.25	2.343	1.700	1.009	47.8	0.791	0.787	0.090†
XGB	−3.64	2.359	1.710	1.016	50.1	0.822	0.591	0.085†

Model	R^2_{OS} (%)	RMSE	MAE	U2	DA (%)	CW p	DM p (Holm)	MCS p
KNN	−1.64	2.336	1.695	1.007	51.1	0.339	1.000	0.275
SVR	−1.29	2.333	1.697	1.008	51.0	0.744	1.000	0.275
DNN	−1.46	2.334	1.700	1.008	49.2	0.386	1.000	0.275
LSTM	−2.54	2.347	1.706	1.016	50.4	0.666	0.353	0.275
Panel B: h = 5								
RW	0.00	5.278	3.832	1.000	.	.	.	1.000
ARX	−9.04	5.512	3.999	1.043	49.6	0.368	0.646	0.181
ENET	−2.44	5.342	3.883	1.016	50.1	0.646	1.000	0.481
ET	−2.09	5.333	3.872	1.013	48.4	0.718	1.000	0.481
XGB	−10.79	5.556	4.062	1.052	47.0	0.785	0.082	0.074†
KNN	−2.35	5.340	3.889	1.016	49.9	0.439	1.000	0.481
SVR	−2.13	5.334	3.870	1.016	50.0	0.615	1.000	0.481
DNN	−4.60	5.398	3.936	1.029	47.4	0.839	0.618	0.173
LSTM	−4.36	5.392	3.910	1.027	49.8	0.918	0.369	0.154
Panel C: h = 22								
RW	0.00	9.555	7.317	1.000	.	.	.	1.000
ARX	−33.10	11.023	8.411	1.184	43.4	0.740	0.226	0.067†
ENET	−16.14	10.297	7.720	1.097	49.7	0.629	1.000	0.156
ET	−12.33	10.127	7.600	1.077	54.9	0.726	1.000	0.156
XGB	−19.82	10.459	7.767	1.107	52.8	0.706	0.661	0.147
KNN	−15.17	10.254	7.715	1.094	52.3	0.647	1.000	0.156
SVR	−18.13	10.385	7.810	1.114	50.6	0.620	1.000	0.150
DNN	−16.41	10.309	7.611	1.104	55.9	0.502	1.000	0.156
LSTM	−27.35	10.783	7.816	1.176	55.8	0.646	1.000	0.156

Notes: Evaluation period 3 January 2022 to 1 October 2026 (1,218, 1,214 and 1,197 forecasts for h = 1, 5 and 22). R^2_{OS}: OOS R^2 against the RW, Eq. [3]. RMSE and MAE refer to h-day log returns in percent. U2: ratio of the price RMSE to that of the RW (RW price RMSE: €1.78, €4.06 and €7.34). DA: share of correctly predicted signs. CW p: one-sided p-value of the Clark and West test against the RW. DM p (Holm): two-sided p-value of the Diebold and Mariano test (HLN version) against the RW, adjusted with the Holm procedure over all 36 pairwise comparisons of the horizon. MCS p: model confidence set p-value (range statistic, 2,000 block bootstrap replications); † marks models outside the 90 percent confidence set.

The formal tests point the same way. The Clark and West statistic, which is designed to give the larger model a fair chance against a nested benchmark, never rejects the null of equal accuracy; its smallest p-value across all 24 model and horizon combinations is 0.34. Directional accuracy hovers around 50 percent. The single exception, the DNN at the monthly horizon with 55.9 percent correct signs, passes the Pesaran and Timmermann test at the 5 percent level, but that test assumes independent forecasts and the 22-day targets overlap heavily, so this result carries little weight. The RW belongs to the confidence set at every horizon with a p-value of one, which means that no model outperforms it in a statistically detectable way. The set excludes the ARX model, ET and XGB at the daily horizon, XGB at the weekly horizon and the ARX model for h = 22.

Pairwise Diebold and Mariano tests sharpen the picture. Before correction, several learners lose significantly to the RW at the 5 percent level, notably at h = 1. After the Holm adjustment across the 36 comparisons per horizon, none of these differences survives at 5 percent; the ARX model at h = 1 and XGB at h = 5 remain worse at 10 percent. No pair of learning models can be told apart

after the adjustment. In other words, the data separate the models from one another far less clearly than a ranking by point estimates would suggest, and the evidence indicates that all of them lose to a forecast of no change by margins that range from negligible to large.

Fig. 3 shows when the losses build up. At the daily horizon, the ARX model, XGB and the LSTM lose most of their ground in 2022, the year of the energy crisis, and then drift sideways. At the monthly horizon all models fall behind sharply early in 2022, when the EUA fell by almost 40 percent in the two weeks after the invasion of Ukraine, and again during the gradual decline of 2023. None of the curves turns upward for long.

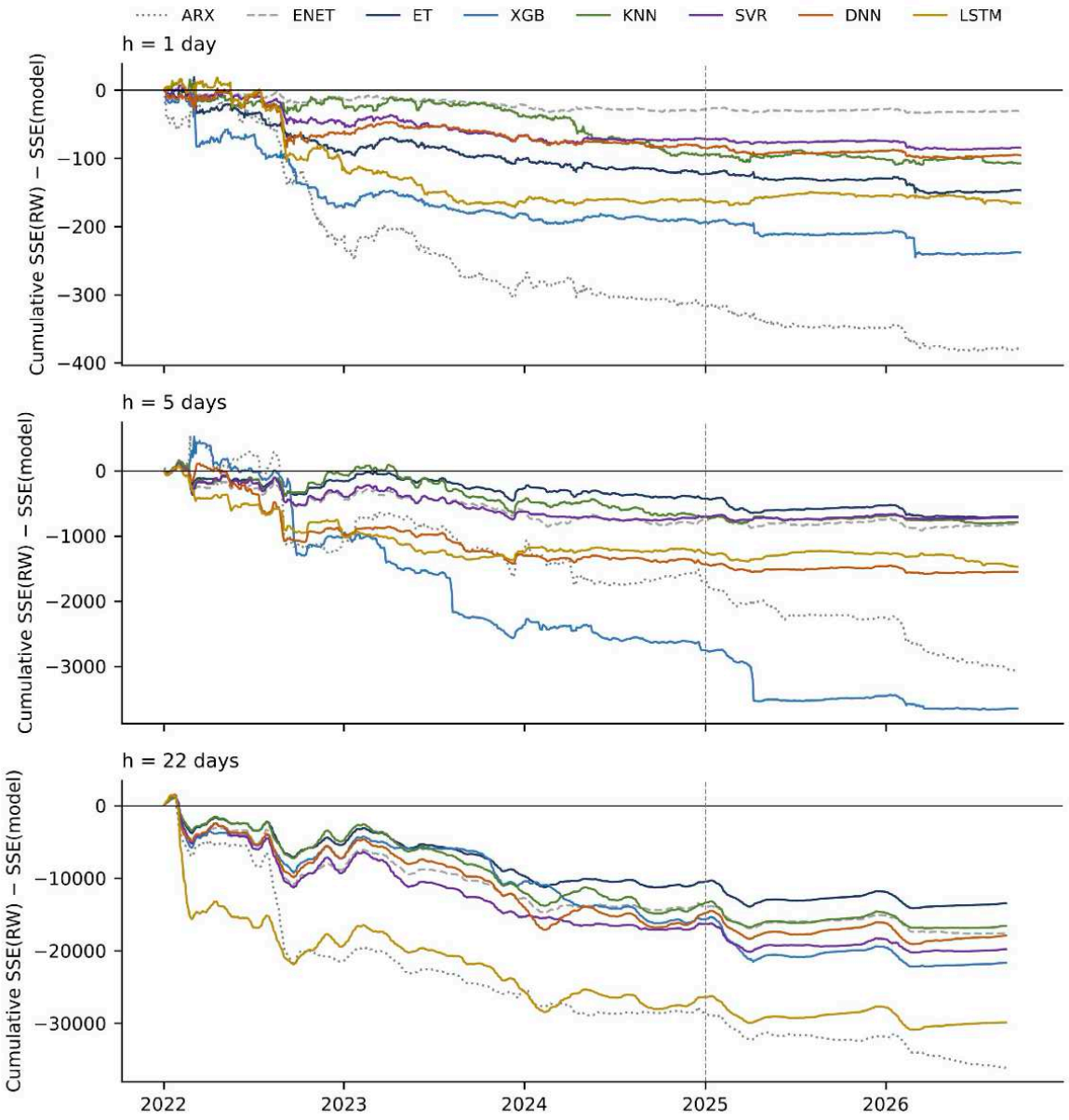


Fig. 3. Cumulative difference between the squared forecast errors of the RW and of every model. A falling line means that the model loses ground to the RW. The dashed vertical line separates sub-periods P1 and P2.

6.2. Where the losses come from

A large part of the gap at longer horizons is unrelated to the predictors. From 2018 to 2021 the allowance price climbed from about €8 to about €80. Any model estimated on that window learns a strong positive drift, and the drift does not repeat after 2021. Treating the historical mean of past returns as the forecast makes this visible (Table 4). Its average forecast is 0.14 percent per day, 0.69 percent per week and 2.99 percent per month, equivalent to an annualised drift of roughly 34 percent, while the realised mean return over the evaluation period is essentially zero. The historical mean therefore loses to the RW by 0.47, 2.24 and 12.05 percent of the squared error at the three horizons.

Table 4. Average forecasts and accuracy against the historical mean.

Forecast	Mean (h=1)	R^2 vs HM (h=1)	Mean (h=5)	R^2 vs HM (h=5)	Mean (h=22)	R^2 vs HM (h=22)
Realised mean return	0.00	.	0.00	.	0.02	.
Historical mean (HM)	0.14	−0.47[a]	0.69	−2.24[a]	2.99	−12.05[a]
ARX	0.03	−5.31	0.15	−6.65	1.99	−18.78
ENET	0.14	0.00	0.69	−0.20	2.37	−3.65
ET	0.07	−1.77	0.40	0.15	2.42	−0.25
XGB	0.11	−3.16	0.52	−8.36	2.40	−6.93
KNN	0.06	−1.16	0.51	−0.11	2.93	−2.78
SVR	0.13	−0.82	0.49	0.12	2.70	−5.42
DNN	0.14	−0.98	0.40	−2.30	3.51	−3.89
LSTM	0.03	−2.05	0.22	−2.07	3.66	−13.65

Notes: Mean: average forecast (or realised value) of the h-day return in percent over the full evaluation period. R^2 vs HM: OOS R^2 in percent with the expanding historical mean as benchmark; positive values indicate that the model beats the historical mean. [a] For the historical mean itself, the entry is its R^2_{OS} against the RW.

Measured against the historical mean instead of the RW, the learners look far more alike. For h = 5, ET and SVR beat the historical mean narrowly (by 0.15 and 0.12 percent), and at the monthly horizon ET loses only 0.25 percent to it. The other models lose more than the drift alone explains, most of all the LSTM and the ARX model. The tuned elastic net shrinks all slope coefficients to zero at the daily horizon and becomes identical to the historical mean; at longer horizons it keeps a few small coefficients and loses slightly more. The lesson is that model flexibility neither recovers the drift error nor adds much information beyond it.

6.3. Sub-periods

Table 5 splits the evaluation period. Losses are smaller in P2 (2025 to 2026) than in P1 at the daily horizon, where the LSTM and the elastic net come within 0.25 percent of the RW and most models stay inside the confidence set with high p-values. At the weekly horizon, SVR loses only 0.34 percent in P2, while the ARX model and XGB do much worse than in P1. At the monthly horizon, losses are of similar size in both sub-periods. The year-by-year results in Appendix Table A2 show

the occasional positive $R^2_{OS,}$ for example the LSTM in 2025 at the daily horizon (1.3 percent) and SVR in 2025 at the weekly horizon (0.8 percent), but these episodes are short and are not shared across models. The calmer conditions of 2025 and 2026 narrow the gap to the RW; they do not reverse it.

Table 5. OOS R^2 by sub-period (%).

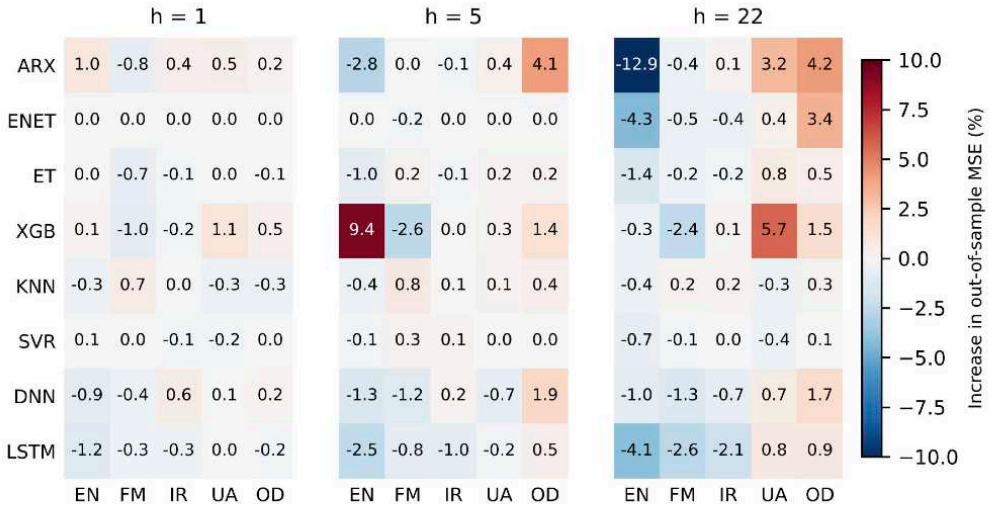
Model	h=1, P1	h=1, P2	h=5, P1	h=5, P2	h=22, P1	h=22, P2
ARX	−6.03	−4.88†	−6.04	−24.67†	−33.35	−32.16†
ENET	−0.53	−0.25	−2.66	−1.31	−16.22	−15.86
ET	−2.34	−1.86	−1.52	−5.06	−12.26	−12.57
XGB	−3.68	−3.48	−9.72	−16.40	−18.26	−25.50
KNN	−1.78	−1.05	−2.51	−1.53	−15.65	−13.45
SVR	−1.37	−0.97	−2.47	−0.34	−19.05	−14.78
DNN	−1.62	−0.82	−5.07	−2.17	−17.30	−13.15
LSTM	−3.12	−0.15	−4.46	−3.83	−30.80	−14.82

Notes: OOS R^2 in percent against the RW. P1: January 2022 to December 2024 (770 forecasts at each horizon); P2: January 2025 to October 2026 (448, 444 and 427 forecasts). † marks models outside the 90 percent model confidence set of the sub-period.

6.4. Predictive relevance of blocks and variables

Since no model beats the benchmark, the relevance scores must be read with care. A positive score means that a model forecasts worse when the variable is scrambled, so the model uses the variable in a way that helps in the test sample. A negative score means that the model forecasts better without it, which is a sign of fitting noise. Across the 21 variables and eight learning models, only 41, 37 and 49 percent of the scores are positive at the three horizons. This share is close to what would be expected if most predictors carried no stable signal.

Fig. 4 reports the scores for the five blocks. Three patterns stand out. First, the own dynamics of the EUA, that is, its recent returns, raise OOS accuracy in all eight models at the weekly and monthly horizons. The gains are small, between 0.01 and 4.2 percent of the MSE, but the consistency across very different learners is notable. Second, the uncertainty and activity block helps six of the eight models at the monthly horizon, by up to 5.7 percent for XGB. Third, the energy block, which theory places at the centre of allowance pricing, makes the forecasts worse in all eight models at the monthly horizon. XGB at the weekly horizon is the exception: scrambling the energy block raises its error by 9.4 percent.



EN: energy fundamentals; FM: financial markets; IR: interest rates; UA: uncertainty and activity; OD: own dynamics. Colours truncated at ±10.

Fig. 4. OOS predictive relevance of predictor blocks: percentage increase in the OOS MSE when all predictors of a block are permuted (Eq. [6]). Positive values indicate that the block improves forecasts; negative values indicate that the model forecasts better without it.

Table 6 explains the third pattern. At the monthly horizon, coal is among the most useful predictors, scoring positively in all eight models; gas does so in seven. The losses come from the renewable share and the switching gap, which rank last and harm every model or nearly every model. Both are level variables whose range shifted out of the training sample during the evaluation period: the renewable share kept rising with capacity additions, and the switching gap reached extreme values in 2022 when the switching price climbed far above the EUA. A relation learned on 2018 to 2021 levels cannot generalise once those levels are left behind. Among the remaining variables, policy uncertainty ranks first at the monthly horizon and second at the weekly horizon, with positive scores in seven and six models respectively.

Table 6. OOS predictive relevance of individual variables.

Rank	h=1	ΔMSE	Pos.	h=5	ΔMSE	Pos.	h=22	ΔMSE	Pos.
1	BRENT	0.08	5/8	EUA	1.08	8/8	EPU	1.06	7/8
2	EPU	0.15	5/8	EPU	0.34	6/8	EUA	1.56	8/8
3	COAL	0.05	5/8	RANGE	0.06	3/8	COAL	0.50	8/8
4	RES	0.21	5/8	UTIL	−0.17	5/8	TTF	0.25	7/8
5	EUA	0.04	4/8	DE10Y	0.02	5/8	VOL	0.28	6/8
6	DE3M	0.01	3/8	VIX	−0.05	3/8	STOXX	0.17	5/8
7	USSPR	0.07	4/8	DE3M	−0.05	4/8	BRENT	0.05	5/8
8	DE10Y	0.04	5/8	NCE	−0.04	3/8	RANGE	0.51	4/8
...									
20	VIX	−0.17	2/8	RES	0.91	2/8	GAP	−2.23	1/8
21	GAP	−0.23	1/8	STOXX	−0.29	1/8	RES	−1.55	0/8
Kendall τ, models		0.01			0.16			0.35	
Kendall τ, P1 vs P2		0.09			−0.43			−0.14	

Notes: Ranking by Borda points computed from the OOS permutation scores (Eq. [6]) of eight learners; ranks 9 to 19 are omitted. ΔMSE: average increase in OOS MSE in percent when the variable is permuted. Pos.: count of models whose score is positive. Kendall τ, models: average rank correlation between the variable rankings of pairs of models. Kendall τ, P1 vs P2: rank correlation between the average scores in the two sub-periods.

The rankings are not stable. The average Kendall rank correlation between the variable rankings of two models is 0.01 at the daily horizon, 0.16 at the weekly horizon and 0.35 at the monthly horizon. Between P1 and P2, the correlation of the average scores is 0.09, −0.43 and −0.14. At short horizons, therefore, models disagree on what matters, and what mattered in the crisis years does not carry over to the calm years. Only own dynamics and policy uncertainty keep a high rank across horizons.

Finally, Table 7 compares OOS relevance with the in-sample importance that SHAP values attribute to the same tree ensembles. In-sample, the leading predictors are the renewable share, past EUA returns, gas returns and the geopolitical risk index, and the signs look economically sensible: a higher renewable share lowers predicted returns, as it should if renewables reduce expected emissions. Yet the rank correlation between in-sample importance and OOS relevance ranges from −0.11 to 0.32. Treating the SHAP rankings as evidence of price drivers would therefore lead to conclusions that the OOS evidence does not support.

Table 7. In-sample importance versus OOS relevance.

h	Model	Top three in sample (SHAP)	Top three in the test sample	Kendall τ
1	ET	EUA, TTF, COAL	RANGE, RES, DE10Y	0.07
1	XGB	TTF, EUA, RANGE	RANGE, EUA, RES	−0.07
5	ET	RES, EUA, GPR	EUA, EPU, EURUSD	0.24
5	XGB	RES, EUA, STOXX	RES, EUA, TTF	0.32
22	ET	GPR, RES, EUA	EPU, EUA, COAL	−0.11
22	XGB	GAP, EUA, GPR	RANGE, EPU, EUA	0.12

Notes: In-sample importance is the mean absolute SHAP value summed over the transformations of a variable, for models fitted on data up to December 2024. OOS relevance is the permutation score of Eq. [6] over the full evaluation period. Kendall τ is the rank correlation of the two orderings across the 21 variables.

6.5. Robustness

Table 8 reports three variations of the design. A rolling window of 750 days, which lets the models forget the 2018 to 2021 drift faster, does not help on balance. It trims the losses of a few models at the daily and weekly horizons, but every model loses more at the monthly horizon, because the shorter window raises estimation noise more than it reduces the drift bias. A reduced predictor set, with the twelve variables that a permutation test on the 2018 to 2021 window found most useful, substantially improves the ARX model at the monthly horizon (from −33.1 to −14.0 percent) and narrows the losses of KNN and ET at the daily horizon, but leaves every R^2_{OS} negative. An equal-weighted combination of the six machine learning forecasts is about as good as the best single

learner at the daily horizon and lies in the middle of its components at longer horizons. It does not beat the RW either.

Table 8. Robustness checks and the effect of same-day predictors.

Specification	ARX	ENET	ET	XGB	KNN	SVR	DNN	LSTM	COMB
Panel A: alternative designs									
h=1: Baseline	−5.81	−0.47	−2.25	−3.64	−1.64	−1.29	−1.46	−2.54	−1.27
h=1: Rolling window (750 days)	−7.34	−0.47	−2.52	−4.18	−3.24	−1.22	−0.76	−2.12	−1.52
h=1: Reduced predictor set	−4.38	−0.47	−0.64	−3.84	−0.55	−1.20	−0.99	−2.18	−1.02
h=5: Baseline	−9.04	−2.44	−2.09	−10.79	−2.35	−2.13	−4.60	−4.36	−2.86
h=5: Rolling window (750 days)	−12.90	−2.48	−2.39	−10.39	−3.78	−3.00	−3.45	−4.69	−3.16
h=5: Reduced predictor set	−4.70	−2.43	−2.37	−11.96	−3.75	−2.97	−5.06	−5.39	−3.84
h=22: Baseline	−33.10	−16.14	−12.33	−19.82	−15.17	−18.13	−16.41	−27.35	−15.84
h=22: Rolling window (750 days)	−52.15	−18.74	−12.77	−28.13	−15.49	−21.54	−21.61	−31.34	−18.86
h=22: Reduced predictor set	−13.96	−11.55	−12.22	−14.88	−15.55	−13.39	−14.42	−12.33	−12.21
Panel B: predictors measured on the target day (h = 1)									
Baseline (predictors dated t)	−5.81	−0.47	−2.25	−3.64	−1.64	−1.29	−1.46	−2.54	−1.27
Same-day predictors	.	1.86	9.59	5.96	3.39	5.01	4.09	.	.

Notes: OOS R^2 in percent against the RW over the full evaluation period. Rolling window: training sample restricted to the last 750 trading days. Reduced predictor set: the twelve variables with the largest permutation scores for an Extra Trees model estimated on the first 80 percent of the 2018 to 2021 window and evaluated on the remaining 20 percent. COMB: equal-weighted average of the six machine learning forecasts. Panel B shifts all predictors except the EUA's own variables and the switching gap forward by one day, so that each is observed on the day of the target return; it serves only to illustrate the effect of leaking future information.

6.6. How much does information timing matter?

To show what is at stake in the timing of predictors, the daily models were re-estimated with one change: the predictors are dated to the day of the target return instead of the day before, and the EUA's own variables are dropped. This mimics a design in which fuel, power and asset prices observed at the close of day t + 1 are used to "forecast" the EUA return of day t + 1. The OOS R^2 jumps from negative values to between 1.9 percent (elastic net) and 9.6 percent (ET), a swing of about twelve percentage points for ET (bottom panel of Table 8). The improvement is real in the sense that EUA returns co-move with returns on gas, power and equities dated the same day. This is not a forecast, because those same-day returns are not available when the forecast would have to be made.

A related pitfall concerns error measures on price levels. The RW alone achieves an average absolute error of 1.68 percent of the price at the daily horizon, 3.85 percent at the weekly horizon and 7.41 percent at the monthly horizon over the evaluation period. Low percentage errors on EUA price levels therefore say very little about forecasting skill unless they are reported next to the RW.

7. Discussion

7.1. What the evidence says about the EUA market

The central finding is that, under a design that uses only information available at the forecast origin, none of the eight models forecasts EUA returns better than a random walk at horizons from one day to one month, and none of the differences in their favour, where they exist in sub-periods, is large or persistent. This is the outcome that a weak-form efficient market predicts (18), in line with the long record of the RW in exchange rate and equity forecasting (37, 38). The EUA market has deepened considerably since 2018, with more participants, more liquidity and an active futures curve, and the results suggest that public information on fuels, power, equities, rates and uncertainty is absorbed into the price within the day.

This reading is reinforced by the contrast between Sections 6.4 and 6.6. The economic channels of Section 2 are visible in the data. EUA returns move with same-day gas, power and equity returns, and in-sample models assign sensible signs to the renewable share and to fuel price changes. What the data do not show is that these channels can be exploited ahead of time. Information about fuel switching appears to reach the allowance price on the day it arrives, not over the following week.

Two partial exceptions deserve attention. The first is the EUA's own past returns, which improve all eight models at weekly and monthly horizons. In the tree ensembles, the SHAP contributions of past returns are negatively correlated with their values, so the models learn a mild tendency for large moves to reverse. Such reversals accord with liquidity-driven price pressure, for example from compliance buyers or from funds adjusting positions, which studies of financial trading in this market have documented (1). The second is the policy uncertainty index, which helps most models at the monthly horizon and enters with a positive sign: periods of high uncertainty tend to be followed by higher returns. This is compatible with a risk premium that compensates holders of allowances for policy risk, which is plausible in a market whose supply depends on legislative decisions (27, 28). Neither effect is strong enough to overturn the RW, and both should be read as associations rather than causal drivers.

7.2. Why fundamentals do not help forecast

The poor OOS showing of the energy block, despite its central role in theory, has two sources. The first is speed: the fuel-switching channel operates through prices that are traded in liquid markets alongside the EUA, so news about gas and coal reaches all three prices at once. The second is non-stationarity of the state variables. The renewable share and the switching gap are exactly the

variables that should matter for allowance prices in the medium run, and in-sample they rank near the top. On test data they rank at the bottom, because their levels moved into territory that the training data did not cover: renewables kept growing, and the 2022 gas shock pushed the switching price to several times the allowance price. Flexible learners, and trees above all, cannot extrapolate such shifts. Returns on coal and on gas, which are stationary, behave much better and appear among the most useful predictors at the monthly horizon.

This suggests a practical lesson for modelling fundamentals: structural state variables need to be expressed in a form that remains comparable across regimes, for instance as deviations from a slowly moving trend, before they are handed to a learner. It also cautions against reading in-sample feature rankings as evidence on price formation, as the weak link between SHAP rankings and OOS relevance shows.

7.3. Implications

For market participants, the results argue for caution about point forecasts of allowance prices. A hedger who takes today's futures quote as the forecast loses nothing against the models tested here, and avoids the drift error that models estimated on a trending sample carry into the future. Effort is better spent on risk measures, scenarios and volatility modelling, where the persistence documented in many studies (26, 31) offers more scope.

For policymakers, the finding that public fundamentals are priced quickly is reassuring about the informational efficiency of the market. It is also a reminder that the MSR and other supply rules work through expectations: if the price responds to fundamentals within the day, policy announcements and their credibility matter more than the slow adjustment of fundamentals. The positive association between policy uncertainty and subsequent returns points the same way. Clear and stable rules on future supply should lower the premium that market participants require for holding allowances.

For researchers, the results support minimum standards for forecasting studies of carbon prices: predictors dated strictly before the forecast origin, with explicit publication lags; tuning and the choice of predictors confined to data that precede the evaluation window; a purge between training targets and the forecast origin when targets span several days; a random walk benchmark, a historical mean benchmark and tests for nested models; and multiple-testing corrections when many models are compared. Many large accuracy gains reported in the literature may not survive these standards. Some may, and replications under a common protocol would help to distinguish the two cases.

7.4. Limitations

The study has limits that point to further work. The analysis uses daily data and horizons up to one month, and its conclusions do not extend to intraday trading or to annual horizons, where policy reforms and the long-run cap dominate. The predictor set is broad but not exhaustive. Positioning data from the weekly Commitments of Traders reports and auction results were left out because they are harder to obtain on a consistent daily basis, and they may carry information that the activity measures used here miss. The model set is compact by design; decomposition hybrids and large pre-trained models were not tested, although the same protocol could be applied to them. Finally, the evaluation window contains the largest energy shock in recent European history. A longer calm period might show more predictability, and the sub-period results hint that the gap to the RW narrows after 2024.

8. Conclusions

This paper set out to establish what level of predictability in EU carbon allowance prices survives a strict forecasting protocol and which economic channels, if any, carry that predictability. EUA futures returns were forecast at one, five and twenty-two trading days with nine models and nineteen daily series grouped by economic channel, using only information available at the forecast origin and tests that account for nested models and multiple comparisons.

On the first question, very little predictability survives. No model beats the RW over January 2022 to October 2026, losses grow with the horizon, and a large share of the long-horizon losses comes from the drift that models learn on the 2018 to 2021 rally. The answer to the second question is more nuanced. Past EUA returns and policy uncertainty contain a small amount of information that helps every model or most models at weekly and monthly horizons. Fuel prices help at the monthly horizon when they enter as returns. Structural state variables, including the renewable share and the fuel-switching gap, dominate in-sample importance rankings yet harm OOS forecasts because their levels shift across regimes. And a design that lets same-day information slip into the predictors would have turned the negative daily results into positive ones.

Future work can build on these results in three directions: by testing whether positioning and auction data add information beyond the measures used here, by modelling the volatility and the distribution of EUA returns rather than their mean, and by applying the same protocol to the hybrid architectures that dominate the recent literature.

Declarations:

Declaration of competing interest:

No competing interest to declare.

Funding:

This research did not receive any specific grant.

Data availability:

All series are publicly available upon reasonable request from the author.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process:

During the preparation of this work the author used a generative AI tool for reshaping visuals, editing, and proof-reading

Appendix A. Hyperparameters and supplementary results

Table A1 lists the search grids and the values selected on the February 2018 to December 2021 window for each horizon. The elastic net selects the largest penalty in the grid at every horizon. A larger penalty would only shrink the coefficients further towards zero and move the forecast closer to the historical mean, so extending the grid would not change the conclusions. Table A2 reports the OOS R^2 by calendar year.

Table A1. Hyperparameter grids and selected values.

Model	Search grid	h = 1	h = 5	h = 22
ENET	penalty $\alpha \in \{0.01, 0.03, 0.1, 0.3, 1\}$; L1 ratio $\in \{0.1, 0.5, 0.9\}$	$\alpha = 1.0$, L1 = 0.5	$\alpha = 1.0$, L1 = 0.9	$\alpha = 1.0$, L1 = 0.1
ET	500 trees; min. leaf size $\in \{10, 30, 60\}$; share of predictors per split $\in \{0.33, 1\}$	leaf 10, share 0.33	leaf 60, share 0.33	leaf 60, share 0.33
XGB	depth $\in \{2, 3\}$; trees $\in \{150, 400\}$; learning rate 0.02; subsample 0.8	depth 2, 150 trees	depth 2, 150 trees	depth 2, 150 trees
KNN	k $\in \{200, 100, 50, 20\}$; uniform or distance weights	k = 50, uniform	k = 100, uniform	k = 100, uniform
SVR	C $\in \{10, 1, 0.1\}$; $\gamma \in \{\text{scale}, 0.005\}$; $\varepsilon \in \{0.1, 0.5\}$	C = 0.1, γ = 0.005, ε = 0.1	C = 0.1, γ = scale, ε = 0.5	C = 1.0, γ = scale, ε = 0.5
DNN	layers (64, 32) or (32, 16); dropout $\in \{0.1, 0.3\}$; weight decay $\in \{10^{-4}, 10^{-3}\}$	(64, 32), drop 0.3, wd 0.001	(32, 16), drop 0.1, wd 0.0001	(64, 32), drop 0.3, wd 0.001
LSTM	hidden units $\in \{16, 32\}$; dropout $\in \{0.1, 0.3\}$; weight decay $\in \{10^{-4}, 10^{-3}\}$; sequence 10 days	16 units, drop 0.3, wd 0.001	16 units, drop 0.1, wd 0.0001	32 units, drop 0.3, wd 0.001

Notes: Selection uses cross-validation over four expanding folds in time order with a gap of h days (ENET, ET, XGB, KNN, SVR) or a hold-out of the last 20 percent of the window (DNN, LSTM), based on February 2018 to December 2021 data only. Networks: Adam optimiser, learning rate 0.001, batch size 64, early stopping with patience 12, forecasts averaged over three seeds. The ARX model has no tuning parameters.

Table A2. OOS R^2 by calendar year (%).

h	Year	ARX	ENET	ET	XGB	KNN	SVR	DNN	LSTM
h=1	2022	−7.8	−0.4	−3.4	−6.2	−0.8	−1.7	−2.2	−4.0
h=1	2023	−5.4	−0.9	−0.7	−0.9	−1.4	−1.7	−0.5	−4.8
h=1	2024	−2.8	−0.4	−1.6	−0.8	−4.2	−0.4	−1.3	0.1
h=1	2025	−5.2	0.0	−1.1	−2.6	−1.3	−0.4	−0.7	1.3
h=1	2026	−4.6	−0.5	−2.6	−4.3	−0.8	−1.5	−1.0	−1.5
h=5	2022	−7.0	−2.8	−1.1	−6.3	−0.4	−2.6	−6.3	−6.5
h=5	2023	−1.8	−3.1	−1.8	−22.9	−6.4	−3.8	−5.2	−3.9
h=5	2024	−7.6	−1.9	−2.3	−5.6	−3.8	−1.1	−2.0	−0.2
h=5	2025	−19.2	0.1	−3.7	−24.3	0.5	0.8	−1.1	0.0
h=5	2026	−30.8	−2.9	−6.6	−7.6	−3.8	−1.6	−3.4	−8.1
h=22	2022	−48.4	−19.8	−12.1	−16.2	−10.9	−20.7	−16.0	−43.6
h=22	2023	−25.1	−24.3	−26.7	−20.5	−44.3	−37.8	−42.1	−38.5
h=22	2024	−12.2	−4.7	−3.3	−20.3	−5.6	−4.1	−3.6	−3.4
h=22	2025	−27.2	−10.6	−11.7	−32.8	−11.8	−17.7	−11.6	−11.9
h=22	2026	−37.2	−21.2	−13.4	−18.0	−15.1	−11.8	−14.7	−17.8

Notes: OOS R^2 in percent against the RW by calendar year (2026: January to October).

References

[1] S. Quemin, M. Pahle, Financials threaten to undermine the functioning of emissions markets, Nat. Clim. Change 13 (2023) 22-31. https://doi.org/10.1038/s41558-022-01560-w

[2] C. Bergmeir, J.M. Benítez, On the use of cross-validation for time series predictor evaluation, Inf. Sci. 191 (2012) 192-213.

[3] S. Gu, B. Kelly, D. Xiu, Empirical asset pricing via machine learning, Rev. Financ. Stud. 33 (2020) 2223-2273.

[4] J.D. Rubin, A model of intertemporal emission trading, banking, and borrowing, J. Environ. Econ. Manag. 31 (1996) 269-286.

[5] B. Hintermann, Allowance price drivers in the first phase of the EU ETS, J. Environ. Econ. Manag. 59 (2010) 43-56.

[6] M. Friedrich, E.-M. Mauer, M. Pahle, O. Tietjen, From fundamentals to financial assets: the evolution of understanding price formation in the EU ETS, EconStor Preprint 225210, ZBW Leibniz Information Centre for Economics, Kiel and Hamburg, 2020.

[7] T.E. Clark, K.D. West, Approximately normal tests for equal predictive accuracy in nested models, J. Econom. 138 (2007) 291-311.

[8] F.X. Diebold, R.S. Mariano, Comparing predictive accuracy, J. Bus. Econ. Stat. 13 (1995) 253-263.

[9] S. Holm, A simple sequentially rejective multiple test procedure, Scand. J. Stat. 6 (1979) 65-70.

[10] P.R. Hansen, A. Lunde, J.M. Nason, The model confidence set, Econometrica 79 (2011) 453-497.

[11] M. Friedrich, S. Fries, M. Pahle, O. Edenhofer, Understanding the explosive trend in EU ETS prices: fundamentals or speculation?, arXiv preprint arXiv:1906.10572v5, 2020.

[12] S. Quemin, R. Trotignon, Emissions trading with rolling horizons, J. Econ. Dyn. Control 125 (2021) 104099. https://doi.org/10.1016/j.jedc.2021.104099

[13] E.D. Delarue, W.D. D'haeseleer, Price determination of ETS allowances through the switching level of coal and gas in the power sector, Int. J. Energy Res. 31 (2007). https://doi.org/10.1002/er.1292

[14] N. Koch, S. Fuss, G. Grosjean, O. Edenhofer, Causes of the EU ETS price drop: recession, CDM, renewable policies or a bit of everything? New evidence, Energy Policy 73 (2014) 676-685.

[15] J. Chevallier, Carbon futures and macroeconomic risk factors: a view from the EU ETS, Energy Econ. 31 (2009) 614-625.

[16] D. Caldara, M. Iacoviello, Measuring geopolitical risk, Am. Econ. Rev. 112 (2022) 1194-1225. https://doi.org/10.1257/aer.20191823

[17] S.R. Baker, N. Bloom, S.J. Davis, Measuring economic policy uncertainty, Q. J. Econ. 131 (2016) 1593-1636.

[18] E.F. Fama, Efficient capital markets: a review of theory and empirical work, J. Finance 25 (1970) 383-417.

[19] P. Aatola, M. Ollikainen, A. Toppinen, Price determination in the EU ETS market: theory and econometric analysis with market fundamentals, Energy Econ. 36 (2013) 380-395.

[20] C. Fezzi, D.W. Bunn, Structural interactions of European carbon trading and energy markets, J. Energy Mark. 2 (2009) 53-69.

[21] B. Zhu, S. Ye, D. Han, P. Wang, K. He, Y. Wei, R. Xie, A multiscale analysis for carbon price drivers, Energy Econ. 78 (2019) 202-216.

[22] K. Duan, X. Ren, Y. Shi, T. Mishra, C. Yan, The marginal impacts of energy prices on carbon price variations: evidence from a quantile-on-quantile approach, Energy Econ. 95 (2021) 105131. https://doi.org/10.1016/j.eneco.2021.105131

[23] M.M. Ghazani, Analyzing the drivers of CO2 allowance prices in EU ETS under the COVID-19 pandemic: considering MEMD approach with a novel filtering procedure, J. Clean. Prod. 427 (2023) 139043.

[24] X. Tan, K. Sirichand, A. Vivian, X. Wang, How connected is the carbon market to energy and financial markets? A systematic analysis of spillovers and dynamics, Energy Econ. 90 (2020) 104870.

[25] X. Ren, Y. Li, F. Wen, Z. Lu, The interrelationship between the carbon market and the green bonds market: evidence from wavelet quantile-on-quantile method, Technol. Forecast. Soc. Change 179 (2022) 121611.

[26] A. Dutta, Modeling and forecasting the volatility of carbon emission market: the role of outliers, time-varying jumps and oil price risk, J. Clean. Prod. 172 (2018) 2773-2781.

[27] G. Perino, M. Willner, Procrastinating reform: the impact of the market stability reserve on the EU ETS, J. Environ. Econ. Manag. 80 (2016) 37-52. https://doi.org/10.1016/j.jeem.2016.09.006

[28] K. Bruninx, M. Ovaere, E. Delarue, The long-term impact of the market stability reserve on the EU emission trading system, Energy Econ. 89 (2020) 104746.

[29] O. Nadirgil, Dynamic transmissions between green and technology stocks, ETFs, commodities, crypto and fiat currencies throughout the global turbulences, J. Clean. Prod. 494 (2025) 145002. https://doi.org/10.1016/j.jclepro.2025.145002

[30] X. Gong, M. Li, K. Guan, C. Sun, Climate change attention and carbon futures return prediction, J. Futures Mark. 43 (2023) 1261-1288.

[31] J. Wang, X. Guo, X. Tan, J. Chevallier, F. Ma, Which exogenous driver is informative in forecasting European carbon volatility: bond, commodity, stock or uncertainty?, Energy Econ. 117 (2023) 106419.

[32] A. Ben Mrad, A. Lahiani, S. Mefteh-Wali, N. Mselmi, Forecasting carbon prices: what is the role of technology?, J. Forecast. 44 (2025) 1867-1883. https://doi.org/10.1002/for.3275

[33] Y. Zhao, Y. Huang, Z. Wang, X. Liu, Carbon futures price forecasting based on feature selection, Eng. Appl. Artif. Intell. 135 (2024) 108646.

[34] Y. Duan, J. Zhang, X. Wang, M. Feng, L. Ma, Forecasting carbon price using signal processing technology and extreme gradient boosting optimized by the whale optimization algorithm, Energy Sci. Eng. 12 (2024) 1-17. https://doi.org/10.1002/ese3.1655

[35] H. Wang, Y. Li, RF-LSTM carbon price prediction based on CEEMDAN decomposition and multiscale entropy reconstruction, Sci. Rep. (2026). https://doi.org/10.1038/s41598-026-35085-5

[36] Y. Xu, D. Lien, How do carbon markets interact with energy-intensive sectors? Evidence from price connectedness, Int. Rev. Econ. Finance 98 (2025) 103857.

[37] R.A. Meese, K. Rogoff, Empirical exchange rate models of the seventies: do they fit out of sample?, J. Int. Econ. 14 (1983) 3-24.

[38] I. Welch, A. Goyal, A comprehensive look at the empirical performance of equity premium prediction, Rev. Financ. Stud. 21 (2008) 1455-1508.

[39] J.Y. Campbell, S.B. Thompson, Predicting excess stock returns out of sample: can anything beat the historical average?, Rev. Financ. Stud. 21 (2008) 1509-1531.

[40] European Parliament and Council of the European Union, Directive (EU) 2018/410 of 14 March 2018 amending Directive 2003/87/EC to enhance cost-effective emission reductions and low-carbon investments, and Decision (EU) 2015/1814, Off. J. Eur. Union L 76 (2018) 3-27.

[41] M. Parkinson, The extreme value method for estimating the variance of the rate of return, J. Bus. 53 (1980) 61-65.

[42] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. R. Stat. Soc. Ser. B 67 (2005) 301-320.

[43] P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, Mach. Learn. 63 (2006) 3-42.

[44] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5-32.

[45] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, 2016, pp. 785-794.

[46] T. Cover, P. Hart, Nearest neighbor pattern classification, IEEE Trans. Inf. Theory 13 (1967) 21-27.

[47] K. Hornik, M. Stinchcombe, H. White, Multilayer feedforward networks are universal approximators, Neural Netw. 2 (1989) 359-366.

[48] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (1997) 1735-1780.

[49] H. Drucker, C.J.C. Burges, L. Kaufman, A. Smola, V. Vapnik, Support vector regression machines, in: Advances in Neural Information Processing Systems 9, MIT Press, Cambridge, MA, 1997, pp. 155-161.

[50] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, A. Zeileis, Conditional variable importance for random forests, BMC Bioinform. 9 (2008) 307.

[51] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, in: Proceedings of the 3rd International Conference on Learning Representations (ICLR), San Diego, 2015.

[52] M.H. Pesaran, A. Timmermann, A simple nonparametric test of predictive performance, J. Bus. Econ. Stat. 10 (1992) 461-465.

[53] D. Harvey, S. Leybourne, P. Newbold, Testing the equality of prediction mean squared errors, Int. J. Forecast. 13 (1997) 281-291.

[54] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems 30, Curran Associates, 2017, pp. 4765-4774.