Pontifícia Universidade Católica
do Rio de Janeiro


**Victor Goulart Oreiro**

# Machine Learning Forecasts of EU ETS Carbon Prices with Economic, Financial, and Policy Uncertainty Variables

**Dissertação de Mestrado**

Dissertation presented to the Programa de Pós-Graduação em Administração de Empresas of PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Administração de Empresas.

Orientador: Prof. Marcelo Cabús Klötzle.

Rio de Janeiro
April 2025

Pontifícia Universidade Católica
do Rio de Janeiro

**Victor Goulart Oreiro**

# Machine Learning Forecasts of EU ETS Carbon Prices with Economic, Financial, and Policy Uncertainty Variables

Dissertation presented to the Programa de Pós-graduação em Administração de Empresas of PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Administração de Empresas. Approved by the Examination Committee:

**Prof. Marcelo Cabús Klötzle**
Advisor
Departamento de Administração - PUC-Rio

**Prof. Carlos de Lamare Bastian Pinto**
Departamento de Administração - PUC-Rio

**Prof. Peter Fernandes Wanke**
UFRJ

Rio de Janeiro, April 25th, 2025

**Victor Goulart Oreiro**

Holds a Bachelor's degree in Economics. His research interests focus on innovative and impactful applications in: Machine Learning and Deep Learning applied to Finance, Economics and Business Administration; Natural Language Processing (NLP); Econometrics; Credit Risk Analysis; Time Series Forecasting; Financial Modeling; Financial Markets; International Economics; Artificial Intelligence; and Large Language Models (LLMs).

Bibliographic data

Oreiro, Victor Goulart

Machine learning forecasts of EU ETS carbon prices with economic, financial, and policy uncertainty variables / Victor Goulart Oreiro ; advisor: Marcelo Cabús Klötzle. – 2025.
75 f. : il. color. ; 30 cm

Dissertação (mestrado)–Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Administração, 2025.
Inclui bibliografia

1. Administração – Teses. 2. Previsão de séries temporais. 3. Incerteza geopolítica. 4. Seleção de variáveis via LASSO. 5. Aprendizado de máquina. 6. Sistema de Comércio de Emissões da União Europeia (EU ETS). I. Klötzle, Marcelo Cabús. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Administração. III. Título.

CDD: 658

## Acknowledgements

Thank you to my advisor for opening the doors to meaningful research themes in finance, for the valuable tips, and for offering support and guidance along the way.

To the professors and staff at IAG, to past professors at other schools, for the knowledge and guidance. Special thanks to Professor Figueiredo for clear and lasting lessons in econometrics, finance, economics, derivatives, and more. Also, a mention to Carlos Bastian, Luiz Brandão, and Jorge Ferreira.

To my family and parents for unconditional support and long-lasting friendship.

To all my classmates, especially to Clara Casartelli, André Ismail and Marcelo Pecly, for supporting and sharing the journey.

Thanks to CAPES for supporting science through financial assistance.

And finally, to the researchers who paved the way for this work, for helping move knowledge and society forward. Your contributions laid the foundation upon which this work stands. I am grateful for the opportunity to add to this continuing dialogue.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

## Abstract

Oreiro, Victor Goulart; Klötzle, Marcelo Cabús. **Machine Learning Forecasts of EU ETS Carbon Prices with Economic, Financial, and Policy Uncertainty Variables**. Rio de Janeiro, 2025. 75 p. Dissertação de Mestrado - Departamento de Administração, Pontifícia Universidade Católica do Rio de Janeiro.

This dissertation analyzes an index representing the price of carbon traded in the European Union Emissions Trading System (EU ETS). By applying a variety of models, traditional econometric approaches (ARIMA), Machine Learning (CatBoost and Random Forest), and Deep Learning techniques (LSTM) were explored. The study utilized a comprehensive set of variables, encompassing traditional economic and financial indicators, as well as alternatives measures related to political, economic, and policy uncertainty. To avoid the risk of overfitting and to improve variable selection, the LASSO regularization technique was applied. In addition to selecting variables to reduce dimensionality, LASSO provided insights into the factors influencing carbon price formation. Among the selected uncertainty variables, the UK Economic Policy Uncertainty and the Climate Transition Risk Index (a proxy for perceived climate policy transition risk) stood out, showing relevance in explaining the dynamics of the S&P Carbon Credit EUA Index. Variable selection via LASSO yielded significant performance gains in out-of-sample tests, reducing overfitting and enhancing the models' generalization capabilities. The consistency of the results was confirmed through time series adequate cross-validation and the Diebold-Mariano test, which verified whether there was a statistically significant difference between the performance of the models. The findings highlight the potential of alternative uncertainty indicators and machine learning methods for forecasting environmental asset prices, showing superior predictive performance in several key validation settings compared to the univariate ARIMA model under the metrics, tests, and validation strategies used.

## Keywords

Time series forecasting; Geopolitical uncertainty; Variable selection via LASSO; Machine learning; European Union Emissions Trading System (EU ETS)

## Resumo

Oreiro, Victor Goulart; Klötzle, Marcelo Cabús. **Previsões com Aprendizado de Máquina dos Preços de Carbono do EU ETS com Variáveis de Incerteza Econômica, Financeira e Política**. Rio de Janeiro, 2025. Número de páginas 75. Dissertação de Mestrado - Departamento de Administração, Pontifícia Universidade Católica do Rio de Janeiro.

Esta dissertação analisa um índice do preço do carbono negociado no *EU ETS*. Por meio da aplicação de diversos modelos, foram exploradas abordagens econométricas tradicionais, aprendizado de máquina, e aprendizado profundo. O estudo utilizou um conjunto abrangente de variáveis, incluindo indicadores econômicos e financeiros tradicionais, bem como medidas alternativas relacionadas à incerteza política, econômica e regulatória. Para evitar o risco de sobreajuste e aprimorar a seleção de variáveis, foi aplicada a técnica de regularização LASSO. Além de permitir a redução de dimensionalidade, o LASSO ofereceu *insights* sobre os fatores que influenciam a formação dos preços de carbono. Entre as variáveis de incerteza selecionadas, destacaram-se o índice de Incerteza de Política Econômica do Reino Unido e o Índice de Risco de Transição Climática, ambos com relevância na explicação da dinâmica do índice *S&P Carbon Credit EUA index*. A seleção de variáveis via LASSO resultou em ganhos significativos de desempenho nos testes fora da amostra, reduzindo o sobreajuste e melhorando a capacidade de generalização dos modelos. A consistência dos resultados foi confirmada por meio de validação cruzada para séries temporais e pelo teste de Diebold-Mariano, que verificou a existência de diferenças estatisticamente significativas no desempenho dos modelos. Os resultados evidenciam o potencial de indicadores alternativos de incerteza e de métodos de aprendizado de máquina na previsão de preços de ativos ambientais, apresentando desempenho preditivo superior, em diversos cenários de validação, quando comparado ao modelo univariado ARIMA, segundo as métricas, testes e estratégias de validação utilizados.



## Palavras-chave

Previsão de séries temporais; Incerteza geopolítica; Seleção de variáveis via LASSO; Aprendizado de máquina; Sistema de Comércio de Emissões da União Europeia (EU ETS)

## Table of Contents

# 1 Introduction

As climate change accelerates, the urgency of finding effective tools to support the energy transition has never been greater. Carbon markets have gradually emerged as one of the more promising approaches in this space. They are gaining traction, but not without their share of complications. Among these mechanisms, the European Union Emissions Trading System (EU ETS) stands apart. Its scale, maturity, and global influence have turned it into a key benchmark for emissions pricing across the world.

The EU ETS plays a foundational role in shaping carbon prices, serving as a model for similar frameworks in other regions. Yet, climate governance is becoming increasingly layered, harder to navigate, and more unpredictable. In response, both academic and market communities are turning their attention to tools that can help make sense of carbon price trends and to forecast them.

Carbon credits behave in a way that sets them apart from traditional financial assets. Their prices are shaped by a wide-ranging combination of forces: regulatory decisions, macroeconomic dynamics, energy shocks, political tension, investor behavior, and more. Earlier research often relied on models like AutoRegressive Integrated Moving Average (ARIMA), Vector AutoRegression (VAR), or Generalized Autoregressive Conditional Heteroskedasticity (GARCH) to analyze these trends. These models tend to struggle when faced with abrupt shifts or nonlinear behavior, which carbon prices exhibit often.

In recent years, advances in machine learning have opened up new possibilities. Especially for time series data that defy regular patterns. Models such as Least Absolute Shrinkage and Selection Operator (LASSO), Random Forest, CatBoost, and Long Short-Term Memory (LSTM) recurrent neural networks have shown promise. They are better equipped to handle the complexities: high dimensionality, non-linear interactions, and unpredictable shifts; that traditional models cannot well capture. Still, despite the growing interest, there is a surprising lack of consistent and comparative studies applying these techniques specifically to carbon markets.

More specifically, few studies include political or geopolitical uncertainty indicators in their forecasting frameworks, even though these variables often

influence investor sentiment and regulatory behavior. Similarly, rigorous model validation, using techniques like rolling cross-validation or significance testing through the Diebold-Mariano test, is still not widely adopted in this field. This dissertation steps into that gap.

It proposes a structured forecasting approach focused on the S&P Carbon Credit EUA Index. The methodology brings together variable selection through LASSO, a mix of linear and nonlinear forecasting models, and a statistical evaluation framework that includes both out-of-sample $R^2$ metrics and rolling window validations, applied in both expanding and fixed forms. It also tests the statistical significance of performance differences between models using the Diebold-Mariano test.

The central goal is to figure out which forecasting models perform best under conditions of real-world uncertainty. And just as important, to identify which variables, from macroeconomic, financial, energy-related, regulatory, and geopolitical indicators, actually help improve predictive accuracy.

To tackle this, the study relies on a comprehensive dataset containing 63 explanatory variables of various types and frequencies. LASSO regularization is used for predictor selection, followed by the application of five forecasting models: ARIMA, LASSO, Random Forest, CatBoost, and LSTM. These models are evaluated using standard forecast error metrics: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Scaled Error (MASE), and Symmetric Mean Absolute Percentage Error (sMAPE**)**; under cross-validation rolling window conditions.

Finally, the Diebold-Mariano test determines if differences in forecast accuracy between models are statistically meaningful, which adds value to the conclusions.

The rest of this dissertation is organized as follows. Section 2 explores the theoretical background: the economic and regulatory foundation of carbon markets, how geopolitical uncertainty shapes price formation, and a critical review of the forecasting literature, particularly studies on the EU ETS. Section 3 outlines the research methodology, explaining how the dataset was built and processed, how variables were selected, and how forecasting models and validation techniques were implemented. Section 4 presents the results, focusing on model comparisons, variable interpretation, and statistical evaluation. Finally, Section 5 concludes the

study by summarizing its key findings, noting its contributions to the literature, and pointing toward future research opportunities.

## 1.1. Justification

As carbon markets become more deeply embedded in climate policy, they are also drawing increasing attention from researchers trying to understand how they work. Within this growing field, the EU Emissions Trading System (EU ETS) stands out. It is the most mature and institutionally structured emissions pricing mechanism to date, and its influence goes well beyond Europe, affecting investment flows and regulatory frameworks around the world.

With the global push for decarbonization accelerating, understanding what drives carbon credit prices is both timely and critical. These assets do not respond to a single force. Instead, their value is shaped by a web of factors: shifts in the economy, movements in financial markets, changes in the energy landscape, new regulations, and episodes of political tension. These influences do not stay permanent. They move, collide and evolve. They often do so in ways that are not linear or easy to predict.

Because of that, traditional forecasting models tend to not perform so well. Structural changes, like the introduction of the Market Stability Reserve (MSR) or the more recent Carbon Border Adjustment Mechanism (CBAM), create sudden shifts, disrupt patterns, and test the limits of standard modeling. These are not just policy adjustments. They represent turning points that change how the market behaves.

In recent years, machine learning and neural network approaches have made significant progress in a wide range of fields, especially in finance and energy. They have shown they can handle, to some extent, noise, nonlinearity, and complexity. Yet when it comes to carbon markets, their use is still surprisingly narrow. Most studies either apply one or two models without appropriate comparison, or they skip key ingredients like robustness checks and statistical tests.

Even more rare are works that bring together a diverse set of forecasting approaches and evaluate them systematically. Especially while also accounting for qualitative dimensions, like uncertainty, investor sentiment, or geopolitical risk,

that cannot be easily measured but clearly shape market expectations as this study demonstrate.

It sets out to build a forecasting framework that is reliable, testable, and above all, useful. A model setup that does not just chase accuracy but also helps make sense of a complex, shifting market. The aim is to offer tools that researchers can replicate, that policymakers can trust, and that investors can use to make better-informed decisions.

## 1.2. General Objective

This research sets out to estimate, assess, and compare the ability of different forecasting models to predict price movements in the S&P Carbon Credit EUA Index. These models are not applied in isolation. More than just benchmarking models, the study aims to uncover which forecasting approaches perform best when exposed to the kinds of complexity and unpredictability that define carbon markets.

## 1.3. Specific Objectives

i. To develop a multivariate dataset that brings together variables from across macroeconomic, financial, energy, regulatory, and geopolitical uncertainty domains. All of them aligned to a monthly frequency. LASSO regression will be applied to sort through the noise and isolate the predictors that matter most.

ii. To compare the forecasting performance of different models using rolling cross-validation and standard error metrics. Differences in model accuracy will also be tested for statistical significance through the Diebold-Mariano test.

iii. To identify the core explanatory drivers behind price movements in the S&P Carbon Credit EUA Index. This involves not just ranking variables but also reflecting on what their predictive power implies for policy, regulation, and market behavior.

# 2 Theoretical Framework

The dissertation is anchored in three foundational pillars. First, the economic and institutional structure of the EU ETS. Second, the impact of geopolitical and macroeconomic uncertainty on the pricing of environmental assets. And third, the main forecasting approaches applied to time series within carbon markets.

Together, these three elements provide a broad yet detailed framework for understanding not only the economic forces that shape carbon pricing, but also the technical and conceptual challenges that come with trying to model them.

## 2.1. The Carbon Market and the EU ETS

Growing international decarbonization commitments has encouraged many countries and economic blocs to adopt carbon pricing instruments. These mechanisms aim to factor the social cost of greenhouse gas (GHG) emissions directly into market decisions, pushing emitters toward more efficient reductions.

The core idea behind carbon markets is relatively simple. Emitters either reduce their emissions or pay to offset them by purchasing allowances (Santikarn et al., 2021; Narassimhan et al., 2018).

Among the available instruments, Emissions Trading Systems (ETS) have taken center stage. These systems place a ceiling on total emissions and allocate or auction tradable allowances to sectors under regulation. The EU ETS, launched in 2005, is the most advanced version of this model. It is widely viewed as the gold standard for emissions trading globally, both for its design and its influence (Ellerman and Buchner, 2007; Santikarn et al., 2021).

Today, the EU ETS covers roughly 40% of total emissions across the European Union (EU). It spans several high-emission sectors, including energy and heavy industry, and more recently, has extended its reach to aviation.

### 2.1.1. Evolution and Phases of the EU ETS

Since its inception, the EU ETS has evolved through a series of regulatory phases, each one marked by structural changes that shaped its current form. Phase I (2005–2007) served mainly as a pilot. It was characterized by an overallocation

of allowances and had little effect on actual emissions reductions. Still, it was a crucial step, a period of institutional learning and experimentation (Ellerman and Buchner, 2007).

Phase II (2008–2012) aligned with the first commitment period of the Kyoto Protocol. It came with tighter rules and a more refined design, yet it continued to face challenges, particularly in the form of persistent oversupply and unstable prices (Chevallier, 2011).

Phase III (2013–2020) marked a turning point. It brought deeper reforms, including a centralized system for allocating allowances, expanded sector coverage, and a transition from free distribution to auction-based mechanisms (Verde et al., 2019).

Phase IV (2021–2030) represents the most substantial leap forward. New stability mechanisms have been introduced, along with long-term regulatory signals designed to address price instability and rebuild market confidence. Among the standout reforms are the MSR, which reshaped how supply is managed (Perino, 2018), and the CBAM, crafted to tackle carbon leakage while aligning trade and climate objectives (European Commission, 2021; Branger et al., 2016).

### 2.1.2. Adjustment Mechanisms and Recent Reforms

In effect since 2019, the MSR was built to bring more balance to the system. It works automatically, pulling surplus allowances out of the market to help reduce excess volatility and promote price stability. Beyond its technical design, its influence on expectations has been notable. Research shows that the MSR has significantly changed how prices are formed, making them more responsive to forecasts about future supply and regulatory actions (Perino and Willner, 2016; Perino, 2018).

Meanwhile, the CBAM was designed to discourage companies from shifting operations to countries with weaker climate policies. Though still in the early stages of implementation, CBAM already signals a shift in how cross-border carbon policy is treated, and it may play a growing role in shaping price behavior and market expectations (European Commission, 2021).

Together with broader European climate initiatives, like the "Fit for 55" package (European Commission, 2021), these reforms have reshaped the character

of the EU ETS. Today, allowance prices no longer reflect just energy supply and demand or basic market fundamentals. They are also deeply tied to political signals and institutional decisions that influence market behavior (Koch et al., 2014).

### 2.1.3. Pricing Instruments and Market Dynamics in the EU ETS

Within the EU ETS, the primary asset traded is the European Union Allowance (EUA), which grants the right to emit one metric ton of $CO_2$ equivalent ($tCO_2e$). EUA prices are determined on secondary markets through auctions and over-the-counter (OTC) trades, and are influenced by a wide set of factors, ranging from regulatory expectations and sectoral demand to external shocks like weather conditions, energy prices, and economic growth (Benz and Trück, 2009; Hintermann et al., 2016; Koch et al., 2014).

Literature has consistently shown that carbon prices in this environment behave differently from traditional financial assets. The market often experiences high volatility, abrupt regime changes, and exogenous shocks driven by regulatory reforms or geopolitical events (Chevallier, 2011; Oberndorfer, 2009). These characteristics make it difficult for conventional econometric models to perform well, as they often rely on assumptions of linearity, structural stability, and ergodic behavior (Han et al., 2019).

Beyond the usual supply and demand variables, more recent research highlights the growing influence of non-economic factors, such as political risk perception, regulatory uncertainty, and overall market sentiment, on carbon pricing (Pastor and Veronesi, 2012; Baker et al., 2016; Caldara and Iacoviello, 2022). As a result, recent models increasingly include qualitative and institutional indicators to improve predictive accuracy.

## 2.2. Geopolitical and Macroeconomic Uncertainty

Carbon markets are subject to a high degree of institutional, regulatory, and political instability. They are not only shaped by economic cycles and energy price shocks, but also answer to radical shifts in environmental governance, unexpected legal rulings, and geopolitical tensions, all of which can reshape how market participants assess risk (Chevallier, 2011).

This kind of uncertainty is especially evident within the EU ETS. The system's regulatory complexity, combined with its central role in the EU's climate framework, amplifies its sensitivity to external shocks. Empirical studies confirm that carbon credit prices are not just about market demand or energy fundamentals. They are also influenced by institutional expectations and political or economic sentiment (Benz and Trück, 2009; Oberndorfer, 2009; Pastor and Veronesi, 2012).

Several real-world events have demonstrated this vulnerability. The United States (U.S.) decision to withdraw from the Paris Agreement in 2017, the energy supply crisis that destabilized Europe in 2021, the consequences of the war in Ukraine, and the EU's ongoing reforms to its emissions market, especially the implementation of mechanisms like the CBAM. All of these have left marks on how carbon prices behave.

Caldara and Iacoviello (2022) argue that geopolitical shocks can have lasting consequences for investment flows and asset valuations. This is particularly true for sectors that are exposed to political and regulatory risk, like environmental markets.

Given this context, it is no surprise that more recent forecasting models have begun to include variables that capture uncertainty and perceptions of risk. Baker et al. (2016) show that economic policy uncertainty meaningfully influences market behavior. Bringing this kind of variable into forecasting models adds nuance, it helps capture market reactions during periods of disruption or regulatory change, when systems are under stress.

### 2.2.1. Sources of Instability and Institutional Dynamics

Carbon markets, especially regulated ones like the EU ETS, are highly sensitive to institutional changes, geopolitical disruptions, and shifts in climate policy. These factors make allowance prices unstable and harder to predict (Oberndorfer, 2009; Chevallier, 2011).

The literature recognizes that exogenous events, such as energy crises, pandemics, court rulings, or even national elections, can affect how economic agents perceive risk. In the case of the EU ETS, structural reforms, especially the implementation of the MSR, have influenced not only the expected supply and demand of the allowances but also increased uncertainty of the regulatory trajectory

(Perino, 2018). Other recent instruments, such as the CBAM and the Fit for 55 package, contribute further to this uncertainty.

Geopolitical factors, such as the war in Ukraine, international sanctions, or trade disputes between major economies, also influence energy prices, and by extension, the carbon market itself (Caldara and Iacoviello, 2022; Wang et al., 2025).

The legal uncertainty surrounding climate litigation and the reluctance of some countries to commit to long-term emission targets are also seen as reasons for uncertainty in climate governance (Battiston et al., 2017; Dai et al., 2022).

These factors make carbon markets quite different from conventional financial markets. Prices do not move just because of market fundamentals, but also based on how people understand new rules and regulations. Pastor and Veronesi (2012) point out that pricing assets linked to public policy requires models that incorporate not only economic fundamentals but also expectations around government action.

This is why recent research increasingly uses numeric indicators to measure institutional and political uncertainty. Beyond improving empirical control, these indicators enhance the performance of forecasting models applied to regulated markets (Battiston et al., 2017; Dai et al., 2022), as the next sections will explore.

### 2.2.2. Uncertainty Indicators and Risk Sentiment

To incorporate institutional, economic, and geopolitical uncertainty in forecasting models, several indices have been developed in recent years based on automated text analysis of newspapers, policy documents, and expert reports. These indicators condense large volumes of media coverage into structured time series and are now widely used in empirical research (Baker et al., 2016; Caldara and Iacoviello, 2022).

The measurement of economic uncertainty has become an essential tool for understanding the effects of public policy on markets. One of the most common indicators is the Economic Policy Uncertainty Index (EPU), developed by Baker et al. (2016), which counts how often newspapers mention terms linked to economic and policy uncertainty.

Building on the EPU, several other indicators have been introduced to capture specific dimensions of uncertainty:

- Overall Equity Market Volatility (EMV Overall) Tracker: This is the main index created by Baker et al. (2019). It measures stock market volatility based on how often newspapers mention terms related to the economy, markets, and uncertainty. The EMV tracker moves closely with the VIX and helps explain what kind of news is driving market volatility. It also serves as the base for more specific indexes, like those focused on macro news or monetary policy.
- Macroeconomic News and Outlook EMV Tracker: measures uncertainty based on news related to GDP, inflation, employment, and financial markets (Baker et al., 2019).
- The Policy-Related EMV Tracker (Policy-Related EMV): captures the portion of stock market volatility driven by political events and discourse, including fiscal and monetary policy, regulation, and national security. It helps identify which political factors are shaping market fluctuations, especially during periods of heightened uncertainty (Baker et al., 2019).
- Energy and Environmental Regulation EMV Tracker (EMV EER): focuses on uncertainty in energy and climate-related regulation, using the same approach (Baker et al., 2019).
- EMV Tracker: Infectious Disease: developed to measure the economic impact of the COVID-19 pandemic (Baker et al., 2019).
- Monetary Policy Uncertainty Index (MPU): captures uncertainty related to central bank actions, with versions based on news (MPU Word News) and academic papers (MPU 10 papers) (Baker et al., 2016).
- US–China Trade Tension Index (UCT): measures the intensity of trade conflict between the world's two largest economies, developed by Rogers et al. (2024).
- Geopolitical Risk Index (GPR): tracks global geopolitical tensions based on the frequency of words related to war, military threats, terrorism, and instability (Caldara and Iacoviello, 2022).
- Global Economic Policy Uncertainty (GEPU): measures worldwide policy uncertainty, with variants adjusted for purchasing power parity (PPP) and real-time data (Davis, 2016).

- Economic Uncertainty Related Queries (EURQ): developed by Bontempi et al. (2021), this index uses Google search volume related to uncertainty in the U.S. and Italy.
- UK Policy Uncertainty Index (UK EPU): a version for United Kingdom (UK) of the EPU adapted for the UK context (using local newspapers) (Baker et al. 2016).
- Climate Policy Uncertainty Index (CPU): focuses specifically on climate and energy policy, based on the EPU structure (Gavriilidis, 2020).
- Energy-Related Uncertainty Indexes (EUI): created by Dang et al. (2023), these measure energy-related uncertainty in 28 countries using data from the Economist Intelligence Unit.
- Climate Risk Index (CRI): combines the Physical Risk Index (PRI) and Transition Risk Index (TRI) to assess the exposure of European financial markets to physical and transition-related climate risks (Bua et al., 2024).
- Twitter Economic Uncertainty Index (TEU) and Twitter Market Uncertainty Index (TMU), created by Baker et al. (2021), use machine learning and real-time Twitter data to track changes in public sentiment and how people perceive economic or market risks. There are different versions of the index: TEU-USA, based on tweets from users in the US; TEU-WGT, which gives more weight to tweets that get more retweets; and TEU-ENG, which includes all tweets written in English (regardless of the location). These versions help capture uncertainty from different points of view, in a quicker and more flexible way than traditional news sources.

Bringing these indicators into the analysis adds value. They help capture uncertainty from multiple angles which is essential for understanding how shifting policies or global events that impact asset prices. Without them, models risk ignoring the dimensions of influence that shape market behavior.

These indices are becoming more visible in academic work, particularly in studies focused on asset pricing, systemic risk, and return forecasting. That said, their use in environmental markets is still catching on and it is a growing field. Early evidence points in a clear direction: when included in predictive models, these uncertainty measures tend to improve performance, especially during moments of crisis or rapid policy shifts (Zhang et al., 2022).

What makes them so useful is not just their content, but their variety. Their thematic diversity and methodological scope allow them to pick up on dynamics that more conventional variables might miss. And in markets like the EU ETS, where institutional signals, political decisions, and geopolitical tensions all influence heavily on pricing, that kind of range matters. It turns abstract risk into something measurable.

### 2.2.3. Empirical Applications and Predictive Relevance

Researchers in economics and finance have widely adopted political, regulatory, and geopolitical uncertainty indices in recent years. Recent studies show that variables like the GPR, EMV, MPU, TEU, and EPU have significant explanatory power over financial assets, affecting returns, volatility, and portfolio decisions (Pastor and Veronesi, 2012; Caldara and Iacoviello, 2022; Baker et al., 2016).

In traditional financial markets, these indices have already been used to forecast stock returns, credit spreads, interest rates, and implied volatility. Pastor and Veronesi (2012), for example, show that uncertainty around economic policy affects risk premia and the valuation of firms exposed to regulation. Supporting this, Gulen and Ion (2016) also show that rising EPU levels reduce corporate investment in the U.S., showing how uncertainty affects real business decisions. Caldara and Iacoviello (2022) also show that spikes in geopolitical risk negatively affect global markets and raise overall risk aversion.

Although this type of research is well established in finance, it is still new in forecasting models for environmental markets. Zhang et al. (2022) note that geopolitical events and political decisions can significantly alter the behavior of the European carbon market. Dai et al. (2022) argue that adding qualitative proxies improves the predictive accuracy of models applied to EUA prices, especially when markets face changes in regulation or political pressure.

Han et al. (2019) also point out that conventional linear models tend to underperform in uncertain settings, and suggest using more flexible methods, based on statistical learning, that can include external variables and nonlinear dynamics.

In this dissertation, the decision to include uncertainty indices from policyuncertainty.com and the GPR index follows that methodological

recommendation. These indicators are used as explanatory variables across different forecasting models so that their contribution to improving the performance of EUA price predictions can be tested, especially during periods of systemic instability.

By combining statistical rigor with qualitative risk proxies, this empirical strategy aims to contribute to a growing research agenda that sees uncertainty as a central factor in price formation within dynamic, regulated, and geopolitically sensitive environmental markets.

Beyond uncertainty indicators, recent literature has also explored how financial variables and market indicators shape the demand for sustainable assets like carbon credits. For example, Baker et al. (2018) and Fatica et al. (2021) show that investor behavior toward green bonds is closely linked to risk appetite and the cost of capital, both of which directly influence the financing of climate transition projects. These insights are particularly relevant to carbon markets, where the appeal of environmental instruments depends on broader financial and regulatory conditions. For this reason, variables like credit spreads and interest rates are also included to indirectly capture the cost of financing and market sentiment toward green assets.

## 2.3. Previous Works

The literature on carbon price forecasting has grown more diverse as emissions trading systems mature. In this context, the following sections present different approaches developed to address the empirical and structural challenges of this market, organized into seven key areas.

### 2.3.1. Traditional Econometric Approaches

A significant part of the literature still relies on classical statistical models, like ARIMA and GARCH, to represent the stochastic behavior of carbon price series (Alberola et al., 2008). In contrast, Chevallier (2011) adopts a more advanced application based on Factor-Augmented VAR (FAVAR) models, incorporating macroeconomic and energy markets shocks to capture structural dynamics often missed by simpler models.

But since these models assume linearity, they do not handle well the kind of abrupt changes often seen in carbon markets. Aatola et al. (2013) show that structural breaks have a strong effect on price patterns in the EU ETS, highlighting the need for approaches that can handle sudden changes in both the regulatory and economic environment. Han et al. (2019) go further by using nonlinear methods to model carbon price volatility, showing the limits of traditional approaches.

### 2.3.2. Hybrid Models, Machine Learning, and Deep Architectures

Because of the limits of traditional models, many studies now use hybrid methods that mix statistical tools with machine learning. Zhu and Chevallier (2017) combined ARIMA with support vector machines and improved forecasting accuracy. Later, Ji et al. (2019) expanded this by including Convolutional Neural Network (CNN) and LSTM networks. They found that this approach worked well for predicting carbon prices in unstable markets.

Hou et al. (2022) used LASSO, support vector regression, and other simple machine learning methods to predict $CO_2$ emissions in China. They used sliding-window cross-validation and showed that good feature selection greatly improves accuracy, especially with small datasets and many variables, like in carbon price forecasting.

Huang et al. (2021) built a hybrid model combining GARCH and LSTM to capture complex time patterns in carbon prices. Xu et al. (2020) went further, using machine learning with network theory and Extreme Learning Machines (ELMs) to improve generalization.

Other recent studies have explored attention-based models for forecasting carbon prices. For example, Wu and Du (2024) propose a dual-stream transformer model with cross-attention, achieving better results than LSTM and Gated Recurrent Unit (GRU) in the European carbon market. Jenko and Costa (2024) use the Temporal Fusion Transformer (TFT) to jointly forecast prices and emissions in multivariate energy systems, reporting significant performance gains. This trend is also supported by the work of Lim et al. (2021), who introduced the TFT. By combining recurrent neural networks with attention mechanisms, TFT strikes a balance between accuracy and interpretability, making it particularly effective for handling volatile time series like carbon pricing.

Zhang (2003) proposed a hybrid model that separates linear and nonlinear parts of a time series. The idea is to use ARIMA to deal with the linear patterns first, and then apply a neural network to model the remaining, more irregular behavior. Even though this was first used in other areas, this kind of approach became popular in finance and environmental studies, because it helps to better understand and predict complex movements, like those we see in carbon markets.

More recently, Liu et al. (2024) developed a hybrid GARCH-LSTM model to forecast carbon prices in five regional markets across China. By combining conditional volatility modeling with the ability of neural networks to capture nonlinear and long-term patterns, the model significantly improved forecasting accuracy, especially in the Hubei, Shenzhen, and Shanghai markets.

Deep learning models, especially LSTM networks, are now widely used in financial time series forecasting. In a systematic review covering 2005 to 2019, Sezer et al. (2020) highlight LSTM as the most widely used deep learning architecture in this field, due to its ability to capture nonlinear behavior and long-term dependencies in economic data. Smyl (2020) also presents a hybrid model that combines exponential smoothing with LSTM, which won the M4 international forecasting competition. This result supports the idea that deep learning models, when regularization and complexity control are applied, can offer strong performance in complex financial contexts.

Around the same time, Oreshkin et al. (2019) presented N-BEATS, a fully connected neural network that breaks away from traditional recurrent and convolutional models. It stood out in the M4 competition for its strong performance and also offers interpretable results through basis expansion, which makes it a solid choice for dynamic markets such as the EU ETS.

Even with all these advances, including hybrid approaches, deep architectures, and attention mechanisms, forecasting financial and environmental time series remains challenging. In one of the largest comparative studies in the field, Makridakis et al. (2020, 2022) present the results of the M4 and M5 Competitions, which tested over 100,000 time series using a wide range of forecasting methods, including neural networks and ensemble models. While the M4 focused broadly on statistical and ML methods, the M5 centered specifically on financial time series during COVID-related uncertainty. Both studies show that predictive performance tends to decline sharply for longer horizons or in periods of

structural change, conditions that are quite common in carbon markets. These findings highlight the importance of using strong methodological practices like time-series validation, complexity control, and variable filtering, all of which are incorporated into the framework proposed in this dissertation.

### 2.3.3. Regime-Switching and Time-Varying Parameter Models

Lin and Zhang (2022) highlight something key about the EU ETS: carbon prices tend to respond sharply to institutional and regulatory shifts. When policy frameworks change, the impact reaches beyond surface-level price dynamics, it often disrupts the core relationships that traditional econometric models rely on. These models, which assume a certain degree of stability and linear behavior, start to underperform. They were not built to handle regime shifts or nonlinear fluctuations.

Because of this, researchers have started moving in another direction. More flexible modeling approaches are gaining ground, that do not lock in static relationships, but instead adjust to regime changes. Markov-Switching models are a good example. Hamilton introduced them in 1989, and since then, the have been widely used to detect changes in volatility, especially when the market moves through periods of uncertainty like economic crises or policy shifts. Chevallier (2011), in his work on the European carbon market, observed that prices frequently switch patterns switching between patterns that tend to fade quickly, usually in reaction to unexpected events like policy changes or market disruptions.

Time-varying parameter (TVP) models offer another perspective. Unlike fixed models, they allow the relationship between variables to shift over time. That is crucial in a context like the EU ETS, where new rules or reforms can quickly alter how prices react to inputs. Ellerman and Buchner (2007), as well as Aatola et al. (2013), argue that tracking these structural changes is essential if we want our models to reflect market behavior with any accuracy.

Recent studies that use TVP-VAR setups let the model’s coefficients change over time. Sometimes it is oil or gas. Other times, electricity or even the stock market. These links do not stay fixed. After 2016, when the Paris Agreement started reshaping global policy and mechanisms like the MSR kicked in, those dynamics began shifting even more. Because these relationships do not stay stable for long,

models that can adjust over time are key to tracking how prices behave in a changing policy context (Li et al., 2021).

Overall, regime-switching and time-varying models are valuable tools for analyzing volatile markets. But they need a lot of data, computing power, and fine-tuning, which can limit their use in real-time forecasting.

### 2.3.4. Mixed-Frequency and MIDAS Regression Models

Forecasting usually comes with the problem of mismatch between the frequencies of available data. While financial market prices are typically available in intervals ranging from seconds and minutes to daily or even monthly observations, many important explanatory variables, such as macroeconomic indicators, fiscal data, or uncertainty indexes, are released weekly, monthly, or quarterly. Unfortunately, standard aggregation methods might lead to information loss, poor lag specification, and distortions in predictive results.

To handle this, mixed-frequency is a useful way to work with time-frequency inconsistencies. Among these, the MIDAS (Mixed Data Sampling) model, introduced by Ghysels et al. (2006), is being one of the most widely used. MIDAS allows variables with different frequencies to be included directly in the modelling, using weights to reflect how each lag of the lower-frequency data affects predictions.

In the context of carbon markets, mixed-frequency models have shown promising results. Zhao et al. (2018) applied the MIDAS approach to combine monthly economic data with daily energy indicators when forecasting EU ETS prices. They found that including mixed-frequency variables significantly improved forecasting accuracy.

Similarly, Niu and Liu (2024) used a GJR-GARCH-MIDAS model to estimate the volatility of EUA futures, showing that monthly macroeconomic variables increased the model's robustness in response to market shocks. These studies show the potential of MIDAS models in environmental time series, which often involve unstable patterns.

### 2.3.5. Effects of Energy and Fuel Prices

Previous studies on carbon pricing show that energy prices, particularly fossil fuels, strongly influence the behavior of emission allowances. This is because oil, natural gas, and coal prices directly affect the operating costs of regulated companies, which in turn influences their emission-reduction strategies and demand for carbon credits. Chevallier (2011) found that energy price shocks significantly affect EUA prices, especially when combined with macroeconomic factors. Supporting this, Bredin and Muckley (2011) identified a negative relationship between energy and carbon prices, linking rising energy costs to a drop in the appeal of emissions trading.

Also, energy source substitution often responds to relative price shifts. For example, when natural gas becomes more expensive than coal, there is an incentive to shift toward more carbon-intensive power generation. This increases the demand for EUAs. This mechanism is known as energy arbitrage and is well documented in studies like Benz and Trück (2009) and Creti et al. (2012).

For this reason, most carbon price forecasting models incorporate energy prices, either directly or through related indicators. Common variables are Brent crude, natural gas, thermal coal, and electricity spot prices. Their inclusion significantly improves model accuracy. Dai et al. (2022) and Benz and Trück (2009), for instance, show strong links between energy market shocks and carbon price volatility in the EU ETS.

Global energy markets have also become increasingly relevant for regional carbon prices. Due to rising integration in trade, finance, and climate policy, EUA prices are now more exposed to external shocks like OPEC+ decisions or geopolitical risks (Creti et al., 2012).

In short, including energy prices as explanatory variables is not only statistically helpful, but also conceptually justified, since the energy variables plays a key role in shaping the marginal cost of emissions reduction.

### 2.3.6. Carbon Market Dynamics and Policy Interventions

The EU ETS is a regulatory system that is in constant evolution, shaped by policy decisions that directly influence supply and demand for the allowances. The literature highlights how rule changes, such as revisions in free allocation, cap

adjustments, or the introduction of market stability mechanisms, can significantly affect carbon prices (Ellerman and Buchner, 2007; Aatola et al., 2013).

One example is the MSR, which started operating in 2019 to reduce price volatility by automatically adjusting the number of allowances in circulation. Perino and Willner (2017) found that the MSR helped the system to gain credibility, pushing EUA prices higher and signaling stronger long-term regulatory commitment.

Another important change is the introduction of CBAM in 2023. It was rolled out step-by-step to bring EU climate rules in line with global trade standards. Basically, it taxes imports from countries with weaker environmental rules, which connects trade, politics, and carbon pricing in new ways (European Commission, 2021).

Even when people expect them, policy changes still shape how the market reacts. As Calel (2013) says, emissions trading only works if prices stay steady and trustworthy. If the rules change too fast or are not very clear, it can scare investors away from green technologies.

Some studies show that policy and political factors can affect carbon prices even more than economic ones. Even though not all of them say exactly how much, research by Dai et al. (2022) and Caldara and Iacoviello (2022) makes it clear that uncertainty in politics or the economy can shake up environmental markets. So, including these variables in carbon price forecasts is justified.

### 2.3.7. Market Sentiment and External Influences

In addition to macroeconomic forces, energy prices, and regulatory developments, carbon markets are also shaped by less tangible, but equally impactful factors: investor sentiment and political confidence. These elements, particularly when tied to policy uncertainty, can intensify volatility in markets like the EU ETS. What makes them so influential is their ability to shift expectations even in the absence of observable changes in fundamentals (Pastor and Veronesi, 2012; Caldara and Iacoviello, 2022; Baker et al., 2016).

Recent research has begun to highlight the role of text-based uncertainty indicators in improving predictive models for environmental assets. Zhang et al. (2022), for example, emphasize the growing inclusion of these variables in carbon

price forecasting. Their advantage lies in their ability to detect external shocks and structural breaks, that traditional variables might miss. This is echoed in the work of Baker et al. (2016) and Caldara and Iacoviello (2022), where both provide strong empirical support for incorporating these indicators in markets that are sensitive to political and policy-driven risks.

Some studies have pushed the boundary even further. Wang et al. (2025) investigated how EPU and GPR influence the volatility of European carbon futures. Using a QARDL model, they found that both indicators were significant predictors, though GPR, in particular, had a more persistent impact during periods of instability. In a different context, Liu and Lü (2023) compared the performance of GPR and CPU indexes against conventional macroeconomic variables when modeling the volatility of China's carbon neutrality index. Their results, derived from GARCH-MIDAS and GARCH-RKV-MIDAS frameworks, suggest that uncertainty metrics carry greater explanatory power in such models.

Another contribution comes from Ghani et al. (2024), who assessed the forecasting accuracy of several uncertainty indexes across sectors exposed to climate risk, such as renewables and transportation. Their findings reinforce the idea that GARCH-MIDAS models are well suited to volatile, regulation-sensitive environments like those found in environmental finance.

Interestingly, financial innovation is also entering the equation. Jin et al. (2020) explored the role of green bonds as a stabilizing instrument within carbon markets. Their study shows that green bonds can act as a hedge against EUA price fluctuations, offering an additional financial mechanism that deserves consideration in forecasting frameworks.

## 3 Methodology

This study builds predictive models to estimate carbon price movements in the EU ETS, using machine learning applied to economic and financial time series. The methodological strategy follows four main steps: (i) preprocessing and preparation of variables, (ii) feature selection using LASSO regression, (iii) training and evaluation of forecasting models, and (iv) performance comparison using the Diebold-Mariano test.

The focus is on forecasting the S&P Carbon Credit EUA Index, which is available in weekly frequency and was converted to monthly for this study. The target variable is modeled based on a broad set of explanatory variables, including macroeconomic, monetary, uncertainty, financial, and environmental indicators, all denominated in U.S. dollars.

Because the data is monthly, the modeling had to deal with some challenges, especially regularization and overfitting, since the number of observations was relatively limited. To address this, LASSO regression was used for variable selection, and rolling window validation was applied with clear separation between training and testing periods. We also used formal statistical tests to compare how the models performed.

The modeling process began by transforming all series to ensure stationarity, a key requirement for linear models and to reduce spurious autocorrelation. After that, all variables were standardized using the z-score method. Then we applied LASSO regression with cross-validation to find the most important predictors and drop redundant or noisy variables. This technique is especially suitable for high-dimensional datasets and helps reduce multicollinearity, as shown by Hastie et al. (2009).

Next, we trained five forecasting models: one linear (LASSO), two tree-based (Random Forest and CatBoost), one statistical (ARIMA), and one deep learning model (LSTM). The idea was to compare different modeling approaches and see how each one performs under different scenarios.

Because flexible models like LSTM are prone to overfitting, we applied additional regularization strategies such as dropout, adaptive learning rate scheduling, and internal validation based on time splits. These techniques help

make the training process more stable and reduce model variance, as recommended in deep learning research for time series (Goodfellow et al., 2016; Brownlee, 2018).

We validated the models using a rolling window cross-validation, more appropriate for time series modelling. Two types of rolling windows were tested, expanding and fixed, to check how each model adapts to different information structures. This strategy is particularly appropriate for financial settings, where data accumulates over time and forecasts must be made using only past information, closely simulating real-world decision-making (Tashman, 2000).

We also used the Diebold-Mariano test to verify whether differences in model performance were statistically significant.

This methodology aims to balance predictive accuracy with statistical rigor and practical usefulness. The combination of modern machine learning, complexity control, and rigorous time-based validation provides a solid framework for understanding carbon price behavior.

### 3.1. Data, Sources, and Preprocessing

This study uses monthly data from the S&P Carbon Credit EUA Index, covering October 2014 to October 2022 (94 observations).

The explanatory variables were selected with a broad scope in mind, aiming to reflect the many forces that drive carbon prices. These include macroeconomic and financial indicators, as well as political, geopolitical, regulatory, and environmental influences.

To ensure robustness and credibility, the independent variables were gathered from widely recognized sources: Eurostat, Bloomberg Terminal, Investing.com, Yahoo Finance, and policyuncertainty.com. In total, the model included 63 explanatory variables. These spanned categories such as energy commodity prices (like Oil, Natural Gas, and Coal); macroeconomic indicators (European unemployment, European oil exports, Euro yield curves, European Inflation); financial metrics (Bitcoin, Commodity indexes, DXY, U.S. Treasury yields (10-year), and bond indexes from FTSE); monetary aggregates (M2 for the U.S., China, and Eurozone); central bank balance sheets (FED and ECB); stock indexes (MSCI Europe, S&P500, FTSE100); volatility indicators (OVX, MOVE); and uncertainty

indexes (detailed in Section 2.2.2). Lagged values of the carbon index itself were also included as a predictor.

Since the variables came in daily, weekly, and monthly formats, we standardized daily/weekly by taking the last available value for each month.

Once aligned, the next step was testing for stationarity, a crucial prerequisite for time series modeling. We applied the Augmented Dickey-Fuller (ADF) test and log-differenced the series where needed. Any variable with a p-value above 0.05 was differenced, and when that was not enough, a second difference was applied, following the guidance of Harris and Sollis (2003). The carbon index series itself (dependent variable) was second-differenced to eliminate unit roots.

A methodological note is warranted regarding the Eurozone monetary aggregate M2 (ECB M2). Although its ADF test resulted in a p-value slightly above the conventional 5% threshold (p=0.075), applying a second differencing significantly reduced forecasting performance for LASSO. After additional tests (KPSS and Jarque-Bera) confirmed acceptable stationarity, we decided a single difference was sufficient. This choice involves a pragmatic trade-off, common in empirical applications, between statistical strictness and predictive effectiveness (Harris and Sollis, 2003; Hyndman and Athanasopoulos, 2018).

With stationarity addressed, we moved on to z-score normalization. Each variable had its mean subtracted and was divided by its standard deviation, a standard transformation aligned with Hyndman and Athanasopoulos (2018) and the Scikit-Learn documentation (Pedregosa et al., 2011). This step is especially important for models sensitive to variable scales. Besides improving performance, standardization helps prevent a single feature from dominating the model and tends to speed up training in gradient-based architectures like recurrent neural networks (Goodfellow et al., 2016).

However, in time series modeling, normalization needs extra care to avoid data leakage. This happens when test set statistics accidentally influence the training process. To prevent this, normalization was done using only the training data for each rolling window. In other words, for each split in the time-series cross-validation, we calculated the mean and standard deviation from the training set only, and applied them to both the training and test sets. This ensures no future information leaks into the past and keeps the forecasting process realistic.

This best practice is widely recommended in time-series forecasting (Hyndman and Athanasopoulos, 2018; Brownlee, 2018), and it was strictly followed throughout the out-of-sample testing in this dissertation. Conditional normalization is particularly critical for models like LSTM, which are very sensitive to data scale and time flow.

Finally, before entering the modeling phase, we applied penalized LASSO regression to select the most relevant variables. This helped cut down the number of predictors, reduce collinearity, and filter out noisy variables. The penalty parameter $\lambda$ was chosen using cross-validation, ensuring stable and robust variable selection.

## 3.2. Econometric and Machine Learning Models

To forecast the European carbon index, we tested five models: LASSO regression, Random Forest, CatBoost, LSTM, and ARIMA. These models were chosen to represent a range of methodological approaches, combining traditional statistical tools with modern machine learning and neural network techniques. The goal was to compare how linear vs. nonlinear models capture patterns in the data.

All models were implemented in Python using well-established libraries like scikit-learn (Pedregosa et al., 2011), catboost, statsmodels, and torch, along with custom scripts for validation control, regularization, and model setup. We based the parameter settings on literature recommendations and fine-tuned them using Time Series Split cross-validation, which helps preserve the time structure and prevent data leakage.

### 3.2.1. LASSO Regression

LASSO regression, introduced by Tibshirani (1996), was used here as both a linear prediction model and a tool for automatic variable selection. What makes LASSO different is the $L_1$ penalty it adds to the regression coefficients. This creates a sparse solution, and many coefficients are reduced to zero, so variable selection happens as part of the estimation.

Its cost function is defined as:

$$\widehat{\boldsymbol{\beta}^{\text{lasso}}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \left( y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \qquad (1)$$

Where $y_i$ represents the target variable, $x_{ij}$ are the explanatory variables, $\beta_j$ are the regression coefficients and $\lambda$ is the regularization parameter that controls how strong the penalty is. Larger values of $\lambda$ apply more shrinkage to the coefficients (eventually pushing some of them to zero) and help keep the model more compact and interpretable.

In practice, the model was implemented using the LassoCV function from the scikit-learn library, with cross-validation based on the TimeSeriesSplit method. A grid of 1,000 lambda values was tested, spaced logarithmically across the range $\lambda \in [10^{-4}, 10^{1}]$. This approach helped balance regularization and overfitting control, while also respecting the time structure of the dataset.

Beyond serving as a standalone model, LASSO was also used during preprocessing as a feature selection filter. The final model coefficients were used to decide which variables to keep: all predictors with non-zero coefficients were selected. This reduced set of variables was then passed to alternative versions of the Random Forest, CatBoost, and LSTM models. It helped reduce dimensionality and added another layer of complexity control.

LASSO was chosen as the filtering method because it is robust in high-multicollinearity environments, something typical in economic and financial datasets, where variables often move together (Hastie et al., 2009). In this study, it proved especially effective in removing noisy features and boosting out-of-sample performance in more complex models.

That made it particularly useful in this case, where the goal was to forecast carbon prices with a small sample and a large number of predictors. Because of the high dimensionality and multicollinearity in the data, LASSO worked well as an automatic variable selection method. Its ability to act both as a prediction model and a feature selector made it a natural first step in the analysis (Tibshirani, 1996; Hastie et al., 2009).

### 3.2.2. ARIMA

ARIMA, developed by Box and Jenkins (1970), is a classic tool for modeling univariate time series, especially useful when capturing linear trends and autocorrelation over time. The ARIMA structure includes three main components: autoregressive (AR), moving average (MA), and integration (I). This allows the model to handle both time persistence and non-stationarity in the series.

The general ARIMA(p,d,q) model can be written as:

$$\boldsymbol{\Phi(L)(1-L)^d y_t = \Theta(L)\varepsilon_t} \quad (2)$$

Where $\boldsymbol{\Phi(L)}$ and $\boldsymbol{\Theta(L)}$ are the lag polynomials for the autoregressive and moving average terms, and d is the number of differences applied to make the series stationary.

In this study, the ARIMA model was applied solely to the dependent variable. Results from the ADF test showed that two differences were necessary to achieve stationarity. Based on these results, the ARIMA (0,2,2) specification was selected as the best fit. The choice was guided by the Akaike Information Criterion (AIC), in line with the recommendations from Hyndman and Athanasopoulos (2018), who emphasize AIC's balance between simplicity and predictive performance.

Although ARIMA is known for its straightforward design, it was deliberately included in the model comparison as a classical statistical benchmark. Its role here was not to outperform more sophisticated techniques but to serve as a point of reference. Including it allows us to assess whether the added complexity of machine learning methods actually translates into meaningful gains in forecasting accuracy.

### 3.2.3. Random Forest Regressor

To go beyond the limits of traditional linear models, this study also uses non-parametric approaches that are more suitable for high-dimensional data and complex variable interactions. One of the most popular of these is the Random Forest algorithm, proposed by Breiman (2001). It is a tree-based machine learning method that uses bootstrap aggregation, also known as bagging. Its main advantage is reducing variance and avoiding overfitting, while still performing well with noisy data and many variables.

The prediction of a Random Forest model is simply the average of the predictions made by all **B** decision trees in the ensemble. Each tree is trained on a different random subset of the data. In mathematical terms:

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} T_b(X) \quad (3)$$

Where B is the total number of trees, and $\boldsymbol{T_b(X)}$ is the prediction of the **B-th** tree trained on a different bootstrap sample. This process helps reduce variance and makes the model more robust, especially in situations where collinearity and noise are present in the data.

In this study, the Random Forest model was implemented using 100 trees (n_estimators=100), with a maximum depth of 6 (max_depth=6). We also set min_samples_split=5 and min_samples_leaf=3, based on tuning done during early testing. These hyperparameters were fixed to ensure stability across the rolling window validation process and to keep the results comparable with the other models.

Because it is an ensemble method, Random Forest handled multicollinearity relatively well and remained effective even with predictors that had low individual importance. It showed solid out-of-sample results, especially when combined with LASSO for variable selection, which helped focus the model on the most relevant inputs.

### 3.2.4. CatBoost

CatBoost (short for Categorical Boosting) is a gradient boosting algorithm developed by researchers at Yandex (Prokhorenkova et al., 2018). Unlike XGBoost or LightGBM, CatBoost improves how categorical data is handled and includes built-in mechanisms to reduce overfitting.

Although the dataset used in this study does not include categorical variables, CatBoost still performs well thanks to techniques like ordered boosting and mechanisms to prevent prediction shift. These features make it especially robust when working with small datasets, like monthly time series.

Technically, CatBoost follows the standard gradient boosting approach. Models $\boldsymbol{h_m(x)}$are fitted one after the other to reduce a loss function $\boldsymbol{L(y, \hat{y})}$. At each iteration **m**, the prediction is updated as:

$$\widehat{y^{(m)}} = \widehat{y^{(m-1)}} + \boldsymbol{\eta} \cdot \boldsymbol{h_m(x)} \qquad (4)$$

Where $\boldsymbol{\eta}$ is the learning rate, and $\boldsymbol{h_m(x)}$ is the base learner trained at iteration m.

CatBoost stands out because of how it processes training data internally. The ordered boosting technique avoids information leakage between training examples and limits sequential dependencies, which helps reduce overfitting, especially when the number of observations is small (Prokhorenkova et al., 2018).

In this study, CatBoost was implemented with 300 iterations (iterations=300), a learning rate of 0.01 (learning_rate=0.01), and tree depth of 4 (depth=4). These values were chosen based on preliminary tests using time-based validation, aiming for a model that is simple but effective. We tested CatBoost with both the full variable set and the reduced version selected by LASSO, to evaluate its performance under different levels of complexity.

The model was trained using the CatBoostRegressor function from the catboost library, with verbose=0 to integrate smoothly with the rolling window loop. All variables were standardized before training, following the same preprocessing steps used for the other models and ensuring no data leakage, as explained earlier.

The results showed that CatBoost was stable and effective in capturing nonlinear patterns among economic variables. Its performance improved even more when combined with LASSO-selected features, suggesting that even tree-based models can benefit from prior dimensionality reduction and multicollinearity control, particularly in small samples.

### 3.2.5. LSTM Neural Networks (Long Short-Term Memory)

LSTM, introduced by Hochreiter and Schmidhuber (1997), is a recurrent neural network widely used for time series tasks in finance and economics. Unlike

standard Recurrent Neural Network (RNN), LSTM has memory cells with gates that decide which information to keep, update, or forget over time. This helps the model learn long-term relationships in the data without losing information, which is a common problem in simpler recurrent networks (Goodfellow et al., 2016).

The LSTM cell works based on a set of equations that define how the forget, input, and output gates operate:

$$\boldsymbol{f_t = \sigma\left(W_f[h_{t-1}, x_t] + b_f\right)}$$
$$\boldsymbol{i_t = \sigma(W_i[h_{t-1}, x_t] + b_i)}$$
$$\boldsymbol{\widetilde{C_t} = \tanh(W_C[h_{t-1}, x_t] + b_C)}$$
$$\boldsymbol{C_t = f_t \cdot C_{t-1} + i_t \cdot \widetilde{C_t}}$$
$$\boldsymbol{o_t = \sigma(W_o[h_{t-1}, x_t] + b_o)}$$
$$\boldsymbol{h_t = o_t \cdot \tanh(C_t)} \quad (5)$$

In the LSTM structure, $\boldsymbol{x_t}$ is the input at time **t**, $\boldsymbol{h_t}$ is the hidden state, and $\boldsymbol{C_t}$ is the cell's internal state. The sigmoid function $\boldsymbol{\sigma}$ is used to control what information passes through the gates. The weight matrices **W** and the bias terms **b** are parameters learned during training. The hyperbolic tangent function $tanh$ is used to keep the values between -1 and 1.

In this work, we implemented a robust version of the LSTM model to reduce overfitting issues, condition that is common in neural networks, especially when working with limited data. The architecture used had two stacked LSTM layers (num_layers=2), each with 32 hidden units (hidden_size=32), a dropout of 0.3 between layers, and a fully connected output layer at the end.

The model was built using the PyTorch library (Paszke et al., 2019) and trained with the Adam optimizer (lr=0.005) and MSE as the loss function.

We applied several regularization techniques to stabilize training. First, early stopping monitored validation loss using 10% of the training set. Second, a learning rate scheduler dynamically adjusted the learning rate during training. Third, input standardization was applied using only the training set in each window to prevent data leakage (Brownlee, 2018). Each model was trained for up to 200 epochs, with a patience of 10 epochs without improvement in validation loss.

The model was trained and evaluated using structured time-based validation, including both rolling windows and sequential splits (TimeSeriesSplit). For each

window, point-by-point forecasts were made and evaluated using robust error metrics (MAE, MSE, RMSE, sMAPE, MASE).

As pointed out by Bergmeir and Benítez (2012), this type of model is very sensitive to temporal information leakage, so it's crucial to use validation methods that respect the data order, like the rolling windows used in this study.

We chose LSTM because it can capture nonlinear dynamics and lagged relationships, both common features in economic and financial time series. In highly volatile contexts with many uncertainty factors, like the carbon market, this flexibility is key. The added robustness in the network setup, including regularization, time-based validation, and sequential standardization, played an important role in achieving stable and accurate out-of-sample performance.

### 3.3. Validation Strategies and Performance Evaluation

Once all models were defined and implemented, the next step was to evaluate how well each one performed in forecasting future values. This stage is critical in time series analysis, particularly in finance, where preserving the natural temporal flow of information is essential. Ignoring the order of data can introduce bias and lead to overly optimistic results. To avoid that, this study followed best practices designed to reflect how models would behave in a real-world forecasting scenario.

Research by Bergmeir et al. (2018) and Cerqueira et al. (2020) highlights the advantages of using rolling window validation over standard cross-validation methods when working with time series. These rolling techniques maintain the sequential nature of data, allowing for more realistic out-of-sample evaluation.

Unlike k-fold cross-validation, which randomizes the data and can introduce data leakage, time series models demand validation strategies that respect chronology. As Tashman (2000) and Bergmeir et al. (2018) point out, shuffling observations breaks the time dependency that these models rely on. Because of that, this analysis used two validation setups: one with a fixed-size rolling window and another with an expanding window that incorporates more data at each step.

The choice is not arbitrary. Bergmeir and Benítez (2012) showed that traditional cross-validation often give biased performance estimates for time series models, especially those like neural networks that are sensitive to lagged inputs.

Their work strongly supports rolling window methods for maintaining the integrity of time-based patterns and ensuring fair comparisons across forecasting models.

For each split in the validation process, we applied a comprehensive set of error metrics. These metrics capture different facets of forecasting accuracy. Some focus on scale, others on percentage deviation or average error. They have become standard tools in applied forecasting research, particularly in large-scale benchmark studies and international forecasting competitions (Makridakis et al., 2018; Hyndman and Koehler, 2006).

### 3.3.1. Out-of-Sample $R^2$ with Temporal Validation

The out-of-sample coefficient of determination ($R^2$) was also used to measure how much better a model performs compared to a simple forecast based on the historical average of the target variable. Even though $R^2$ is usually linked to in-sample fit, it can also be used in forecasting tasks, especially with rolling window setups, to measure the added value of a model over a naive baseline (Kuhn and Johnson, 2013; Bergmeir et al., 2018). We used the classic formula implemented through the r2_score function from the scikit-learn library, which compares the model's squared errors to the ones from a constant mean forecast.

The formula used for $R^2$ is:

$$R^2 = 1 - \frac{\sum(y_t - \hat{y}_t)^2}{\sum(y_t - \bar{y})^2} \quad (6)$$

where $\boldsymbol{y_t}$ are the actual values in the test set, $\boldsymbol{\hat{y}_t}$ are the predicted values from the model, and $\bar{y}$ is the average of the actual values. A positive $R^2$ means the model performs better than the naive mean forecast. A negative value means it performed worse.

The out-of-sample $R^2$, which compares model predictions against a naive baseline, should not be the sole indicator of predictive accuracy, as relying exclusively on this comparative metric can be misleading. The literature strongly recommends using a combination of metrics, complemented by formal statistical tests, to obtain a complete and robust evaluation of forecasting performance, particularly in economic and financial applications, where data can be volatile and

prone to abrupt changes (Hyndman and Athanasopoulos, 2018; Bergmeir et al., 2018).

### 3.3.2. Expanding Rolling Window Cross-Validation

In expanding window validation, the training set starts with an initial block of data that gradually increases over time. At each new step, fresh observations are added to the training window, while the test set remains fixed in size. This structure mirrors the way information is processed in real economic contexts, where agents learn progressively, adjusting their expectations as new data becomes available (Tashman, 2000).

Formally, if **t** is the total number of observations and **h** is the size of the test window, the training and testing sets at each iteration **i** are defined as:

$$\text{Train}_i = \{1,2,\dots,t_i\}, \quad \text{Test}_i = \{t_i + 1,\dots,t_i + h\} \tag{7}$$

where $\boldsymbol{t_i}$ increases at each step.

For this study, we applied three expanding windows, each calibrated to ensure a balanced representation across the full sample period. The idea was to make sure that every model had exposure to different market conditions at different stages of training.

Importantly, models were retrained from scratch at each iteration. No information carried over from one window to the next. Just as crucial, input variables were standardized using only the statistics from the current training set, not the full dataset. This step, although sometimes overlooked, is critical for preserving the integrity of time-based validation. As Bergmeir et al. (2018) emphasize, using global statistics can leak future information into the past, which compromises the realism of forecasting results.

By strictly applying this window-by-window approach, we ensured that our performance measures reflect a true out-of-sample setting, one that closely resembles how forecasting would work in practice.

### 3.3.3. Fixed Rolling Window Cross-Validation

The fixed window strategy keeps the training set size constant throughout the iterations. Each time the window moves forward, the oldest data is dropped and the newest data is added, so the number of training observations stays the same. This method is useful when only the most recent information is considered relevant for forecasting, which is often the case in volatile or fast-changing environments like financial and energy markets. Research has shown that this approach can lead to more stable performance when there are regime shifts or structural changes (Tashman, 2000; Cerqueira et al., 2020).

The training and test sets for iteration **i** are defined as:

$$\text{Train}_i = \{t_i - w + 1, \dots, t_i\}, \quad \text{Test}_i = \{t_i + 1, \dots, t_i + h\} \tag{8}$$

where **w** is the fixed size of the training window.

Here as well, three validation blocks were used to match the expanding window setup and allow direct comparison. The models were re-estimated at each step, and the data was standardized using statistics from the training set only.

The fixed window is often recommended when there are signs of structural breaks or regime changes, as it tests how well models adapt to recent market behavior. This is especially important when older data may no longer represent current patterns, a point emphasized by Hyndman and Athanasopoulos (2018) and Cerqueira et al. (2020).

### 3.3.4. Evaluation Metrics

The evaluation metrics used in this study follow well-established practices in time series forecasting literature, as discussed by Hyndman and Koehler (2006). Five different metrics were applied. Each one captures different aspects of model performance, and they work as complementary tools.

All these metrics were applied to the out-of-sample predictions generated under both TimeSeriesSplit and rolling window (fixed and expanding) setups. This approach is in line with recommendations from Hyndman and Athanasopoulos (2018), who suggest evaluating models over distinct time blocks to simulate real-world forecasting situations and avoid overfitting.

The MSE was used as the main metric, since it gives more weight to larger errors and is sensitive to outliers. This makes it a good choice for capturing severe prediction mistakes (Hyndman and Koehler, 2006). It was also used as the base metric in the Diebold-Mariano statistical test to compare models (see section 3.5). Using MSE here helps check whether the performance differences between two models are statistically significant over time (Diebold and Mariano, 1995; Harvey et al., 1997).

$$\mathrm{MSE} = \frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2 \tag{9}$$

The MAE gives a straightforward measure of average prediction error in the same units as the target variable. It is less sensitive to outliers than MSE, making it a solid choice to evaluate absolute accuracy. RMSE is the square root of MSE, which puts the result back on the same scale as the original data. This metric is especially useful when we want to emphasize big errors, and it performs well when prediction errors are roughly normally distributed (Hyndman and Koehler, 2006; Chai and Draxler, 2014). Both metrics are widely used in forecasting studies because they reflect different dimensions of accuracy.

$$\mathrm{MAE} = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t| \tag{10}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2} \tag{11}$$

The sMAPE expresses forecast error in percentage terms and treats over and under predictions symmetrically. It uses both predicted and observed values in the denominator, which helps avoid distortion when the actual values are close to zero. This makes sMAPE useful for comparing models across different time series scales. However, some studies show that sMAPE can behave unpredictably in extreme cases, so it is best used alongside more robust metrics (Hyndman and Koehler, 2006).

$$\text{sMAPE} = \frac{100\%}{n}\sum_{t=1}^{n}\frac{|y_t - \widehat{y_t}|}{(|y_t| + |\widehat{y_t}|)/2} \tag{12}$$

The MASE, introduced by Hyndman and Koehler (2006), was especially useful in this study because it compares model performance to a simple naive forecast, usually defined as $\widehat{y_t} = y_{t-1}$. A MASE value below 1 means the model is doing better than the naive benchmark, which is important in financial time series where simple rules often perform surprisingly well.

$$\text{MASE} = \frac{\frac{1}{n}\sum_{t=1}^{n}|y_t - \widehat{y_t}|}{\frac{1}{n-1}\sum_{t=2}^{n}|y_t - y_{t-1}|} \tag{13}$$

Combining these metrics provides a more complete view of model performance, balancing absolute error, sensitivity to outliers, and comparison to naive baselines. All of them were applied consistently to every out-of-sample validation window, and the average results were calculated for each metric at the end of every cycle. This kind of setup, as recommended by Tashman (2000) and Hyndman and Koehler (2006), ensures that results do not rely on a single measure and supports stronger conclusions about model performance. This full set of metrics was applied systematically across all out-of-sample validation windows.

### 3.4. Diebold-Mariano Test

An essential step in this study was to formally compare the predictive performance of the models. For that, we used the Diebold-Mariano (DM) test, proposed by Diebold and Mariano (1995), which checks if the difference in forecasting errors between two competing models is statistically significant. The DM test is widely used in time series and economic forecasting contexts (Hyndman and Athanasopoulos, 2018).

The DM test starts from the null hypothesis that both models have the same forecasting accuracy, meaning that their expected errors are equal. The test statistic is based on the series of differences between the loss values (usually squared errors) from each model across the forecast horizon. In this study, we used MSE as the loss

function, since it is sensitive to large deviations and commonly used in predictive modeling.

Using the DM test is particularly important in situations like this, where we are dealing with non-stationary time series, possible regime shifts, and structural breaks, all common in the carbon market. In these cases, just comparing average metrics like MSE, RMSE, or MAE may miss key differences at specific points in time (Pesaran and Timmermann, 1995). Because the DM test evaluates the prediction errors step by step, it is known to be more sensitive to those fluctuations.

The DM test is also well-suited for handling serial correlation in forecast errors, a common issue in rolling window validation. To apply the test properly, we used the Newey-West adjustment for estimating variance, with the optimal lag set as int$\left(1.2 \cdot T^{1/3}\right)$, as suggested by Diebold and Mariano (1995). We also included the correction proposed by Harvey et al. (1997), which adjusts the test statistic for small samples and improves accuracy when there is residual autocorrelation. The use of the DM test is well established in the forecasting and time series literature as a formal way to compare predictive models (Diebold and Mariano, 1995; Hyndman and Athanasopoulos, 2018).

The test statistic is calculated as:

$$DM = \frac{\bar{d}}{\sqrt{\frac{2\pi \hat{f}_{\bar{d}}(0)}{T}}} \tag{14}$$

where $\bar{\boldsymbol{d}}$ is the mean difference in forecasting errors between the two models across **T** periods, and $\widehat{\boldsymbol{V}}(\bar{\boldsymbol{d}})$ is a robust estimate of the variance of that mean, adjusted for possible autocorrelation in the residuals.

In forecasting, it is not enough to show that one model has a lower average error. As discussed by Pesaran and Timmermann (1995), it is important to apply tests that check both the statistical and economic relevance of these differences. That is why using the DM test here helps confirm whether the differences in performance are meaningful or just random noise.

We ran pairwise comparisons between all five main models under both modeling setups (with and without variable selection via LASSO). The forecast

errors used were the ones from the rolling window validations, in both expanding and fixed formats, so we could capture performance differences more reliably.

The DM test results, presented in the next chapter, point out which models actually outperformed the rest. This test is especially useful to go beyond just looking at average error metrics and gives statistical evidence when one model is actually better than another.

### 3.5. Limitations

Despite the use of a robust and systematic methodology, a few limitations remain that are worth highlighting:

(i) Variable selection via LASSO: While LASSO plays a valuable role in reducing dimensionality, it can also exclude variables that may be theoretically important, even if they are not strong predictors in a specific window. Moreover, the set of selected variables tends to shift depending on the estimation window, which can affect the internal consistency of the forecasting results.

(ii) Sensitivity of complex models such as LSTM: Recurrent neural networks like LSTM are powerful but computationally demanding. They rely heavily on hyperparameter tuning and large datasets to reach their full potential. Due to computational constraints, we were not able to perform a comprehensive hyperparameter search, which may have limited the LSTM's overall performance in this context (although its great final result).

(iii) Assumptions behind rolling window validation: Using fixed and expanding windows does a good job of replicating real-time forecasting setups. Still, it assumes some level of structural stability across windows. In the case of carbon markets, where external shocks, regulatory reforms like CBAM, and changes to the EU ETS are common, this assumption does not always hold.

(iv) Structural limits in forecasting financial time series: No matter how advanced the model, forecasting carbon prices remains a difficult task. These assets are heavily exposed to political uncertainty, sudden macroeconomic events, and institutional shifts. Such volatility makes long-horizon predictions particularly fragile, regardless of the algorithm used.

(v) Treatment of the ECB M2 variable: One methodological exception involved the ECB M2 monetary aggregate. Although its ADF p-value (0.075) was

slightly above the usual 5%, we opted for a single differencing. This choice was based on the fact that applying a second difference substantially worsened model performance for LASSO. While the decision was supported by the KPSS and Jarque-Bera tests (which confirmed acceptable stationarity and normal distribution after the first difference), it still introduced a small inconsistency in preprocessing.

Beyond these technical aspects, it is also worth noting that the models employed here are predictive in nature, not causal. While LASSO helps enhance predictive power by filtering variables, its selections should not be interpreted as definitive drivers of carbon price movements.

Even with rolling window validation, abrupt disruptions, such as the COVID-19 pandemic or geopolitical crises, can easily distort forecasts. These limitations do not undermine the findings, but they should be kept in mind when interpreting the results or applying the framework to future data.

# 4 Data Analysis

This chapter delivers the empirical results from the forecasting models and validation strategies introduced in Chapter 3. The aim is to systematically compare how different models perform in forecasting the S&P Carbon Credit EUA index.

## 4.1. Descriptive Statistics

Table 1 shows the main descriptive statistics for the variables selected by LASSO. These statistics are commonly used in financial research, along with the results of stationarity tests for the explanatory variables included in the forecasting models. Given the macroeconomic, financial, and uncertainty nature of the dataset, many series required first or even second differencing to achieve stationarity.

Table 1 - Descriptive Statistics and Stationarity Tests (ADF and JB)

| Variable | Mean | Median | SD | Min | Max | Kt | Sk | ADF | JB |
|---|---|---|---|---|---|---|---|---|---|
| S&P Carbon Credit EUA Lag | 0.00 | 0.01 | 0.18 | 0.46 | 0.60 | 4.27 | 0.37 | 0.01 | 0.01 |
| UK EPU | 0.01 | 0.03 | 0.33 | 0.68 | 0.84 | 2.70 | 0.17 | 0.03 | 0.68 |
| European Oil Exports | 0.00 | 0.00 | 0.06 | 0.14 | 0.14 | 2.63 | 0.05 | 0.01 | 0.75 |
| MSCI Europe Index | 0.00 | 0.00 | 0.05 | 0.18 | 0.17 | 5.02 | -0.39 | 0.05 | 0.00 |
| Climate Risk TRI | -0.01 | -0.01 | 0.02 | 0.08 | 0.11 | 9.04 | 1.25 | 0.01 | 0.00 |
| ECB M2 | 0.00 | 0.00 | 0.02 | 0.07 | 0.05 | 3.28 | -0.09 | 0.08 | 0.81 |
| FTSE Interest Rate | 0.00 | 0.00 | 0.02 | 0.07 | 0.06 | 4.13 | -0.24 | 0.01 | 0.05 |
| Policy-Related EMV | 0.00 | 0.04 | 0.35 | 0.82 | 0.87 | 2.58 | -0.20 | 0.01 | 0.52 |
| MPU World News | 0.02 | -0.02 | 0.47 | 0.80 | 1.45 | 3.43 | 0.66 | 0.01 | 0.02 |
| Industry Index (Europe) | 0.00 | 0.00 | 0.01 | 0.04 | 0.04 | 8.02 | -0.50 | 0.01 | 0.00 |
| UCT | 0.00 | 0.00 | 0.18 | 0.48 | 0.52 | 2.99 | 0.16 | 0.01 | 0.83 |
| FTSE High Yield | 0.00 | 0.00 | 0.04 | 0.15 | 0.22 | 13.93 | 1.56 | 0.01 | 0.00 |
| EMV EER | 0.31 | 0.27 | 0.24 | 0.00 | 1.17 | 4.27 | 1.09 | 0.03 | 0.00 |

| Variable | Mean | Median | SD | Min | Max | Kt | Sk | ADF | JB |
|---|---|---|---|---|---|---|---|---|---|
| GPR | 0.01 | 0.00 | 0.22 | 0.60 | 0.62 | 3.70 | 0.21 | 0.01 | 0.27 |
| HICP | 0.00 | 0.00 | 0.01 | 0.02 | 0.02 | 5.59 | -0.06 | 0.01 | 0.00 |
| Bloomberg Commodity Index | 0.00 | 0.00 | 0.04 | 0.12 | 0.14 | 4.27 | -0.12 | 0.02 | 0.04 |
| GSCI Industrial Metals | 0.00 | 0.00 | 0.07 | 0.19 | 0.17 | 2.84 | -0.15 | 0.01 | 0.79 |
| Natural Gas | 0.01 | -0.01 | 0.14 | 0.42 | 0.36 | 3.91 | -0.12 | 0.01 | 0.18 |
| Bitcoin | 0.04 | 0.05 | 0.23 | 0.62 | 0.52 | 3.08 | -0.30 | 0.01 | 0.50 |
| TEU ENG | 0.01 | 0.03 | 0.60 | 2.41 | 2.20 | 6.53 | -0.10 | 0.01 | 0.00 |
| Energy Risk Global | 0.00 | -0.01 | 0.21 | 0.39 | 0.76 | 3.97 | 0.81 | 0.01 | 0.00 |
| EMV Overall | 0.00 | 0.02 | 0.30 | 0.82 | 0.95 | 3.21 | -0.04 | 0.01 | 0.90 |
| TMU WGT | 0.04 | -0.08 | 0.83 | 1.72 | 2.27 | 2.68 | 0.31 | 0.01 | 0.38 |
| FED M2 | 0.00 | -0.01 | 0.09 | 0.26 | 0.38 | 5.95 | 0.53 | 0.01 | 0.00 |

Note: SD = Standard Deviation; Min = Minimum; Max = Maximum; Kt = Kurtosis; Sk = Skewness; ADF = Augmented Dickey-Fuller test p-value; JB = Jarque-Bera test p-value.

The Eurozone monetary aggregate M2 (ECB M2), received a different treatment. Although its ADF p-value was 0.08, a little above the conventional 5% threshold, it was differenced only one time. This decision followed empirical testing: when this variable was included in its second differenced form, the out-of-sample performance of the models worsened, particularly LASSO, which is highly sensitive to the information content of its inputs. This approach aligns with the principle that, in predictive modeling, practical forecasting accuracy can sometimes justify small deviations from strict statistical criteria (Hyndman and Athanasopoulos, 2018; Shmueli, 2010). Since the main goal of this study is forecasting, and not structural inference, the practical predictive gain justified its inclusion at a little more relaxed significance level.

On the other hand, the U.S. monetary aggregate M2 (FED M2), had to be differenced twice, as its first difference still showed a high p-value (0.71) in the ADF test. This behavior is typical of an I(2) process, meaning the series contains two unit roots so it requires second differencing to achieve stationarity. It reflects the presence of a highly persistent stochastic trend that cannot be removed with a single differencing.

While the ADF p-value suggested borderline non-stationarity, other diagnostics reinforced the decision to include ECB M2 in the model. The Jarque-Bera p-value for ECB M2 is 0.81, indicating no significant deviation from normality. The KPSS test statistic for ECB M2 is 0.144, well below the critical values for rejecting the null of stationarity. These results support the conclusion that first differencing was sufficient and that the series probably behaves as stationary in practice (Kwiatkowski et al., 1992; Harris and Sollis, 2003). This multiple testing approach, using both ADF and KPSS, follows recommendations from the time series literature to improve the robustness of stationarity diagnosis. Together with the Jarque-Bera test, this reinforced the statistical adequacy of the first-differenced series.

This cautious approach to variable transformation was consistently applied throughout the dataset. Most explanatory variables were accepted as stationary at the 5% level after one or two rounds of differencing, with exceptions like the one discussed.

In addition to stationarity, Table 1 summarizes key features of the distributional shape of the series. Several variables show skewness and excess kurtosis, common features in financial and economic time series. Yet most pass the Jarque-Bera test, indicating that normality is a reasonable assumption for the majority of inputs. This contributes to model stability, particularly for regularized techniques like LASSO, which tend to perform better with input distributions that are approximately symmetric and not heavy-tailed (Zou and Hastie, 2005; Brownlee, 2018).

Taken together, these diagnostics confirm that the dataset was well-prepared and statistically adequate to modeling.

Since the main goal of this work is forecasting accuracy, not causal inference, using a tighter statistical filter but with some flexibility was the best approach. This pragmatic balance between statistical theory and empirical performance reflects a growing trend in applied forecasting, particularly in volatile domains such as finance and environmental economics (Shmueli, 2010; Hyndman and Athanasopoulos, 2018).

## 4.2. Variable Selection and Economic Interpretation via LASSO

After completing the LASSO-based variable selection outlined in Section 3.2.1, this section presents the final set of variables retained in the model, along with their standardized coefficients.

Before running the estimation, all predictors were standardized using z-scores. This step is essential in LASSO regression, as it ensures the regularization penalty applies equally across variables, independently of their original scale. With the optimal penalty parameter in place, the model identified 24 predictors with non-zero coefficients.

The selected variables cover macroeconomic, financial, energy, regulatory, climate, uncertainty, and geopolitical aspects, as well as lagged values of the carbon price series itself. This composition provides a comprehensive foundation for the predictive model.

Table 2 lists each selected variable and its corresponding standardized coefficient. What stands out is the frequency with which qualitative and institutional indicators appear in the model. These are often difficult to quantify, yet they were consistently retained. So this highlight something important: even in data-driven models, investor sentiment and institutional signals may play a stronger role in shaping price expectations than purely numerical trends would suggest.

Table 2 - LASSO Selected Variables (Standardized Coefficients)

| Explanatory Variable | Standardized Coefficient |
|---|---|
| S&P Carbon Credit EUA Lag | -0.074 |
| UK EPU* | -0.045 |
| European Oil Exports | 0.030 |
| MSCI Europe Index | 0.030 |
| Climate Risk TRI* | -0.026 |
| ECB M2 | 0.024 |
| FTSE Interest Rate | -0.018 |
| Policy-Related EMV* | -0.018 |
| MPU World News* | 0.017 |
| Industry Index (Europe) | -0.016 |
| UCT* | -0.015 |
| FTSE High Yield | 0.015 |
| EMV EER* | -0.012 |
| GPR* | -0.011 |
| HICP | 0.010 |

| Table 2 - LASSO Selected Variables (Standardized Coefficients) | |
|---|---|
| Explanatory Variable | Standardized Coefficient |
| Bloomberg Commodity Index | -0.008 |
| GSCI Industrial Metals | -0.007 |
| Natural Gas | -0.006 |
| Bitcoin | -0.003 |
| TEU ENG* | -0.003 |
| Energy Risk Global* | -0.003 |
| EMV Overall* | -0.003 |
| TMU WGT* | -0.002 |
| FED M2 | 0.002 |
| Note: This table reports variables selected by the LASSO regularization procedure. Coefficients are standardized to allow magnitude comparison across predictors. Asterisks (*) indicate qualitative (text-derived) or institutional variables. | |

While LASSO is primarily used for forecasting, the variables it selects, along with the sign of their coefficients, can still offer valuable insights into the determinants of carbon prices. Since regularization shrinks coefficient magnitudes, they should not be taken as precise measures of impact. Still, a positive coefficient on an uncertainty index, for example, suggests that carbon prices tend to rise as perceived risk increases. Such patterns help build intuition about market dynamics, even though they do not imply causality.

Among the variables tied to environmental policy and transition risk, one stood out: the Climate Risk TRI, which entered the model with a negative coefficient of -0.026. Built from text-based data, TRI captures how uncertain the regulatory environment appears around decarbonization and energy transition policies. That negative sign is consistent with earlier findings. Unexpected policy changes or abrupt shifts in climate regulations tend to amplify risk and supress the value of regulated assets (Battiston et al., 2017).

TRI's inclusion in the model suggests that carbon markets respond negatively when public discourse around ecological transition becomes more intense. This may reflect expectations of future changes in permit allocation, shifts in regulated demand, or revisions to sectoral incentives. This interpretation aligns with both theory and recent empirical evidence. Battiston et al. (2017) found that sudden changes in climate policy can sharply reprice assets, particularly in heavily regulated sectors. Similarly, Bua et al. (2024) show that shocks to the TRI influence

investor behavior, especially in the post-2015 period, affecting green and brown assets differently.

On the macro-financial side, some variables reflecting liquidity and real economic activity were retained. The ECB's monetary aggregate M2 (ECB M2) appeared with a positive coefficient (0.024), suggesting that monetary expansion may drive up carbon credit prices, possibly by boosting demand in regulated sectors or in sustainable assets more broadly. That aligns with work by Fatica et al. (2021), who highlighted liquidity's role in supporting the valuation of green assets. The U.S. monetary aggregate (FED M2), differenced twice, also had a positive (although smaller) effect (0.002). However, it would be misleading to claim ECB M2 had a stronger influence, given the different treatment in transformations applied to each series. But in general, European assets tend to be more representative to the determination of European carbon prices, probably due to market's geographic and institutional scope.

Two indicators supported the idea that stronger economic activity pushes carbon prices higher. The MSCI Europe index and European oil exports were both selected, each with coefficients of 0.030. This fits with prior studies, such as Creti et al. (2012), which linked carbon prices to economic growth and commodity trade activity in the European context.

Energy prices, by contrast, had negative signs. Natural gas futures entered with a coefficient of -0.006, and the Bloomberg commodity index with -0.008. One possible interpretation is that higher input costs lower the demand for carbon credits, either because they make compliance more costly or because they draw political focus away from tightening regulations. This echoes findings from Bredin and Muckley (2011), who observed that energy shocks can weaken carbon pricing, especially in times of policy uncertainty.

The HICP inflation index entered with a moderate positive coefficient (0.010), suggesting that carbon prices may respond to general price dynamics, or may behave like regulated assets that retain value during inflationary periods. Koch et al. (2014) noted a comparable long-term link between inflation and EUA prices.

Political and geopolitical uncertainty indicators also played a significant role. The UK EPU had the largest absolute coefficient (-0.045) among all variables. That is particularly interesting given that the UK left the EU ETS in 2020 (although this study comprises the period from 2014 to 2022). Even so, instability in the UK's

climate policy seems to spread across the region, possibly affecting broader expectations in the European market. Similar regional spillovers have been documented by Zhang et al. (2022) and Benz and Trück (2009).

The Policy-Related EMV was also retained, with a strong negative coefficient, reinforcing that global regulatory uncertainty matters for the pricing of environmental assets. That is consistent with studies by Wang et al. (2025) and Ghani et al. (2024), which suggest a tight link between policy volatility and investor risk premiums.

Geopolitical tension, captured by the GPR, showed a negative coefficient (-0.011). It points to a familiar mechanism: in periods of conflict or diplomatic tension, investors demand a higher premium for long-term or regulated assets, reducing demand in carbon markets (Caldara and Iacoviello, 2022).

Interestingly, indicators based on social media sentiment were also kept. The TEU ENG index came in with a coefficient of -0.003, and TMU WGT slightly lower at -0.002. This suggests that carbon markets are increasingly sensitive not only to traditional indicators, but also to digital sentiment captured by online discourse.

The US-China Trade Tension Index (UCT) also entered with a negative coefficient (-0.015), suggesting that diplomatic tensions and trade barriers between the world's two largest economies can affect negatively European carbon prices. This reflects systemic interdependencies, as also highlighted by Zhang et al. (2022) and Ghani et al. (2024).

Two FTSE fixed-income indexes were selected: the high-yield bond index (0.016) and the interest rate index (-0.018). The high-yield signal suggests that in times of rising credit risk, investors may turn to alternatives, like carbon credits. That is in line with evidence showing a growing shift toward ESG products in riskier markets (Baker et al., 2018; Fatica et al., 2021). The interest rate index reinforces the idea that rising rates, by increasing capital costs, discourage long-term investment, particularly in projects tied to energy transition or environmental goals, which rely heavily on financing and are subject to regulation (Bredin and Muckley, 2011).

Lastly, the model retained the lag of the carbon index itself, with a coefficient of -0.074, the highest absolute value in the set. This highlights a strong autoregressive component, likely reflecting mean-reverting behavior. Regulated

markets often move this way, balancing between policy signals and short-term corrections. Chevallier (2011) observed this pattern, identifying short memory structures in carbon prices, consistent with markets that are regularly rebalanced by regulatory supply-demand adjustments.

In summary, the LASSO model retained a suitable set of variables that influence carbon pricing. By reducing the original 63 variables to 24, the model did not lose predictive power; instead, it became more interpretable and stable in out-of-sample forecasts. While the goal here was predictive performance, the results still shed light on the kinds of forces shaping carbon markets. As noted by Creti et al. (2012), Dai et al. (2022), and Ghani et al. (2024), carbon pricing is not driven by any single factor, but rather by the combined weight of economic fundamentals, political signals, and evolving perceptions of risk.

### 4.3. In-Sample and Out-of-Sample Evaluation

The empirical analysis began by testing how well the proposed models performed under both in-sample and out-of-sample conditions. This initial step served two purposes: spotting early performance trends and identifying signs of overfitting that might undermine predictive reliability. To do this, we ran the models twice, once using the full set of explanatory variables, and again using the reduced subset selected via LASSO. The idea was to compare results and see how variable selection shaped overall forecast accuracy.

While in-sample evaluation is, of course, limited by its reliance on past data, it still offers a useful benchmark for how well each model fits the observed series. Out-of-sample testing, on the other hand, tells a different story. Using time-based cross-validation, specifically TimeSeriesSplit with five folds, gives a more realistic picture of how the models might perform in real-world forecasting.

Table 3 brings together the results for all five models, tested with and without variable selection. A few consistent patterns emerged, but there were also some surprises. As expected, non-linear models generally posted stronger in-sample $R^2$ scores, with LSTM achieving the highest fit in both feature setups. However, this performance did not generalize well when using the full feature set, where the out-of-sample $R^2$ turned negative, a clear sign of overfitting. When combined with LASSO-selected features, however, LSTM's predictive performance improved

substantially, outperforming the other models. Interestingly, CatBoost and Random Forest showed slightly better out-of-sample performance when trained on the LASSO-selected features.

Table 3 - $R^2$ In-Sample and Out-of-Sample Results

| Model | $R^2$ In-sample (Full features) | $R^2$ In-sample (LASSO) | $R^2$ Out-of-sample (Full features) | $R^2$ Out-of-sample (LASSO) |
|---|---|---|---|---|
| LASSO | 0.702 | 0.752 | 0.177 | 0.287 |
| CatBoost | 0.711 | 0.715 | 0.114 | 0.172 |
| Random Forest | 0.835 | 0.836 | 0.131 | 0.190 |
| LSTM | 0.849 | 0.854 | -0.210 | 0.333 |
| ARIMA | 0.474 | 0.474 | 0.095 | 0.095 |

Note: $R^2$ values are shown for in-sample and out-of-sample performance under both full and LASSO-selected feature sets. The out-of-sample $R^2$ measures how well each model predicts unseen data compared to a naive baseline. Higher values indicate better predictive performance.

More meaningful insights come from the out-of-sample validation. LASSO maintains stable performance in both scenarios, showing its ability to generalize. The LASSO model also gained from its own selection process. When it was trained only on the predictors it had selected through regularization (a setup we could call LASSO-LASSO) its out-of-sample performance improved noticeably, with $R^2$ rising from 0.177 to 0.287.

This result for LASSO reinforces the benefits of combining regularization and model parsimony in complex settings. In contrast, ARIMA struggled to maintain accuracy in the out-of-sample results. This supports earlier critiques by Han et al. (2019) and Aatola et al. (2013), who pointed to the model's sensitivity to regime shifts and instability in parameter estimation.

Machine learning models such as CatBoost, Random Forest, and LSTM exhibited a pattern of behavior. While they risk overfitting in their raw form (especially for LSTM with the full set of variables), using LASSO for variable selection was crucial for reducing this risk and, more importantly, improving out-of-sample forecast performance. LSTM, in particular, benefited greatly from the prior LASSO selection, showing significant improvement compared to the version without feature selection. While this pattern was expected due to the model's

sensitivity to high-dimensionality, the degree of improvement was still striking, especially considering the relatively small sample size and high noise typical of financial market data.

This result supports the findings of Sezer et al. (2020), who highlight the potential of LSTM models in high-dimensional settings with complex nonlinearities, as long as they are assisted by proper pre-selection processes. The combination of L1 regularization (LASSO) and recurrent neural networks can, therefore, be seen as an effective hybrid architecture for modeling environmental financial markets, as suggested by Ji et al. (2019) and Zhang (2003) in other economic contexts.

These results highlight the need to evaluate not just average accuracy, but also model consistency and predictive robustness. Good in-sample performance can be misleading if it is not accompanied by generalization ability, especially in series subject to geopolitical shocks and regulatory changes, as is the case with carbon prices. This reinforces the importance of combining multiple performance metrics with time-aware validation tools.

In the next sections, these findings will be explored with more rigorous validation, allowing us to observe how the models behave over time and under forecasting conditions closer to the reality faced by economic agents and regulators.

## 4.4. Evaluation with Rolling Window Cross-Validation

After the earlier in-sample and out-of-sample analysis, this section goes deeper into how robust the models really are performing, using rolling window cross-validation in both expanding and fixed setups. This kind of time-based validation is widely recommended in the economic and financial literature (Pesaran and Timmermann, 1995; Tashman, 2000) because it better reflects how decisions are made in real time. It keeps the order of the data intact and helps avoid data leakage.

### 4.4.1. Expanding Rolling Window

In the expanding window setup, the training window gets larger with each new iteration, adding more recent information as it moves forward. The results,

shown in Table 4, reveal clear performance improvements across all models when LASSO is used to select the most relevant variables.

The biggest improvement came from the LSTM. Its MSE dropped from 0.0277 (0.0049) to 0.0170 (0.0025)**,** and its RMSE went down from 0.1660 (0.0153) to 0.1300 (0.0095). The sMAPE, which is important for measuring percentage errors in a balanced way, also improved a lot, going from 123.522% (11.2156) to 102.775% (9.1329). These results confirm that reducing the number of variables helped the model perform better, more consistently, and with less risk.

Table 4 - Expanding Window Cross-Validation Results

| Model | MSE | MAE | RMSE | sMAPE | MASE | Feature Set |
|---|---|---|---|---|---|---|
| ARIMA | 0.0242 (0.0054) | 0.1216 (0.0103) | 0.1547 (0.0179) | 188.384 (4.3013) | 0.6925 (0.0944) | Full features |
| CatBoost | 0.0238 (0.0070) | 0.1139 (0.0117) | 0.1531 (0.0220) | 131.088 (19.3791) | 0.6475 (0.0864) | |
| LASSO | 0.0239 (0.0076) | 0.1231 (0.0116) | 0.1531 (0.0255) | 122.173 (8.5040) | 0.7024 (0.1130) | |
| LSTM | 0.0277 (0.0049) | 0.1283 (0.0154) | 0.1660 (0.0153) | 123.522 (11.2156) | 0.7254 (0.0478) | |
| Random Forest | 0.0239 (0.0064) | 0.1180 (0.0092) | 0.1538 (0.0200) | 124.546 (18.7176) | 0.6712 (0.0793) | |
| ARIMA | 0.0242 (0.0054) | 0.1226 (0.0103) | 0.1547 (0.0179) | 188.384 (4.3013) | 0.6925 (0.0944) | LASSO Selected Features |
| CatBoost | 0.0226 (0.0072) | 0.1104 (0.0111) | 0.1492 (0.0231) | 126.815 (23.5936) | 0.6278 (0.0869) | |
| LASSO | 0.0182 (0.0058) | 0.1094 (0.0241) | 0.1337 (0.0211) | 109.088 (22.2578) | 0.6226 (0.1513) | |
| LSTM | 0.0170 (0.0025) | 0.0981 (0.0052) | 0.1300 (0.0095) | 102.775 (9.1329) | 0.5584 (0.0603) | |
| Random Forest | 0.0235 (0.0083) | 0.1195 (0.0127) | 0.1518 (0.0261) | 127.833 (21.2241) | 0.6811 (0.1119) | |

Note: Reported values correspond to means with standard deviations in parentheses. Forecasting performance was evaluated using an expanding-window cross-validation approach.

The MASE for LSTM also showed a clear improvement, dropping from 0.7254 (0.0478) to 0.5584 (0.0603). Since it is below 1, this indicates that the model outperformed the naive benchmark.

The LASSO model, when applied using only the variables it had previously selected (in LASSO-selected features), delivered competitive results against the others: an MSE of 0.0182 (0.0058) and an sMAPE of 109.088% (22.2578). This highlights that, despite being a linear model, LASSO can perform competitively. In this case it performed even greater when applied on top of its own selected features.

CatBoost improved slightly, although with a higher standard deviation on some metrics. Random Forest, on the other hand, showed no clear gains and even registered a small drop in some metrics. For example, CatBoost saw its sMAPE drop from 131.088% to 126.815%, although with higher standard deviation (19.3791 to 23.5936); while its MASE went from 0.6475 (0.0864) to 0.6278 (0.0869). This behavior is consistent with the literature, which highlights the natural robustness of ensemble models to noise and multicollinearity (Breiman, 2001; Prokhorenkova et al., 2018).

The ARIMA model, being univariate, was not affected by variable selection and showed weaker results, especially with compared to LASSO-Selected Features models. Its sMAPE for example stayed above 180% (precisely 188.384 (4.3013)), confirming its limitations in capturing the complex dynamics of the carbon market.

### 4.4.2. Fixed Rolling Window

In the fixed window setup, the size of the training set remains constant throughout the evaluation. This approach reflects a forecasting environment where only the most recent data is considered relevant, a useful assumption in markets that are prone to structural breaks or regime shifts, such as the EU ETS.

The results, shown in Table 5, generally confirmed the patterns observed earlier. Once again, the LSTM model using the LASSO-selected variables outperformed all others across every error metric, although with a bit more fluctuation. It posted the lowest MSE at 0.0267 (0.0139), and its RMSE fell to 0.1597 (0.0427). sMAPE came in at 106.6670% (11.4217), the best among all models tested. MASE, too, remained solid, landing at 0.6781 (0.2568).

| Table 5 - Fixed Window Cross-Validation Results | | | | | | |
|---|---|---|---|---|---|---|
| Model | MSE | MAE | RMSE | sMAPE | MASE | Feature Set |
| ARIMA | 0.0421 (-) | 0.1441 (-) | 0.2051 (-) | 195.440 (-) | 0.7205 (-) | Full Features |
| CatBoost | 0.0331 (0.0122) | 0.1307 (0.0196) | 0.1795 (0.0359) | 153.315 (19.7233) | 0.7305 (0.2114) | |
| LASSO | 0.0331 (0.0130) | 0.1319 (0.0233) | 0.1791 (0.0386) | 153.209 (33.4340) | 0.7401 (0.2392) | |
| LSTM | 0.0369 (0.0179) | 0.1397 (0.0333) | 0.1874 (0.0519) | 135.835 (9.8020) | 0.7859 (0.2897) | |
| Random Forest | 0.0345 (0.0144) | 0.1334 (0.0176) | 0.1827 (0.0411) | 141.064 (17.6481) | 0.7444 (0.2003) | |
| ARIMA | 0.0421 (-) | 0.1441 (-) | 0.2051 (-) | 195.440 (-) | 0.7205 (-) | LASSO Selected Features |
| CatBoost | 0.0316 (0.0108) | 0.1274 (0.0159) | 0.1758 (0.0322) | 144.037 (16.0282) | 0.7110 (0.1905) | |
| LASSO | 0.0380 (0.0042) | 0.1475 (0.0114) | 0.1947 (0.0110) | 152.840 (34.0117) | 0.8186 (0.1707) | |
| LSTM | 0.0267 (0.0139) | 0.1200 (0.0257) | 0.1597 (0.0427) | 106.667 (11.4217) | 0.6781 (0.2568) | |
| Random Forest | 0.0330 (0.0132) | 0.1299 (0.0102) | 0.1793 (0.0362) | 134.656 (24.1619) | 0.7210 (0.1485) | |

Note: Reported values are means with standard deviations in parentheses. All models were evaluated using a fixed rolling-window cross-validation scheme. For the ARIMA model, standard deviations are not reported in some cases due to non-convergence in the last two windows. This likely occurred because the maximum likelihood estimation used by ARIMA can fail when data in a window lacks sufficient variation or information to reliably estimate model parameters.

By contrast, LASSO showed signs of instability in this setup. Its MSE increased slightly from 0.0331 to 0.0380, MAE rose from 0.1319 to 0.1475, RMSE moved from 0.1791 to 0.1947, and MASE climbed to 0.8186. That relatively high MASE suggests that LASSO's forecasts, when compared against a naive baseline, were considerably less reliable in this configuration. The model's performance appears to have been more sensitive to the fixed window design, likely due to the reduced training size and an inability to adjust effectively to shifting regimes. Linear models, by nature, often struggle under these conditions. Without the flexibility to adapt quickly, they tend to lose stability and generalization ability when working with shorter, non-expanding datasets.

CatBoost and Random Forest, on the other hand, handled the setup with more consistency. Both models maintained competitive performance, including after the variable selection step. In fact, they showed modest gains in certain metrics. For instance, CatBoost achieved an MSE of 0.0316 (0.0108) and brought its sMAPE down to 144.0370% (16.0282), a noticeable improvement over its full-variable version at 153.3150% (19.7233). This pattern matches what is found in the literature, where these algorithms are known to handle changes in feature sets well (Prokhorenkova et al., 2018).

Meanwhile, ARIMA had the weakest performance across all metrics and showed no change, as it does not react to changes in the explanatory variables.

### 4.4.3. Comparative Takeaways

Looking at the overall picture, some strong conclusions stand out:

(i) In general variable selection using LASSO made a real difference in predictive performance. It was not just a quick fix; it helped reduce variance and made the models more stable overall. This was especially true for LSTM, which tends to be more sensitive to irrelevant inputs (Sezer et al., 2020).

(ii) LSTM, in fact, was the model that reacted most strongly to preprocessing. Initially, it lagged behind in the early tests. But once the right variables were selected, it climbed to the top, outperforming the others in most metrics under rolling window evaluation.

(iii) The true predictive robustness of the models only became clear under sequential validation.

(iv) In the expanding window setup, LASSO performed noticeably better when using the variable set it had previously selected. That is probably not a coincidence. Expanding windows allow models to gradually learn from more data, and it can learn more stable and reliable patterns. And when the noise (those less relevant predictors) is removed in advance, the model starts from a cleaner place. It overfits less in the early stages and refines better as more observations come in. This behavior fits well with what Bergmeir et al. (2018) describe about model stability in time series cross-validation. It also echoes findings from Hou et al. (2022), who show LASSO’s potential when paired with cross-validation strategies in forecasting tasks.

(v) Interestingly, LASSO's performance was more fragile under the fixed window validation. Unlike the expanding setup, where it gained from accumulating more data, here the model struggled to generalize. Its MSE, MAE, and MASE increased, suggesting that linear models like LASSO may be more sensitive to reduced sample sizes and abrupt shifts in data structure. This reinforces the importance of matching model type to validation design, and shows that regularization alone cannot fully compensate for structural limitations.

This analysis highlights how important it is to use multiple validation windows, as recommended by Tashman (2000) and Hyndman and Athanasopoulos (2018). It also shows that, when it comes to forecasting carbon prices, which are clearly nonlinear and unstable, deep learning models combined with regularization techniques offer a promising path forward.

## 4.5. Diebold-Mariano Test

To complement the analysis of model performance, the Diebold-Mariano (DM) test was applied. This statistical test is used to compare the predictive accuracy of two competing models by looking at the differences in their forecast errors over time.

The test was run using both the full set of explanatory variables and the reduced set selected by LASSO, covering all possible pairwise combinations of the models evaluated.

The Diebold-Mariano test results for the main model comparisons are shown in Table 6. The table highlights which performance differences are statistically significant at the 5% level.

Table 6 - Diebold-Mariano Test Results

| Model 1 | Model 2 | Loss Function | DM Statistic | p-value | Feature Set |
|---|---|---|---|---|---|
| LASSO | CatBoost | MSE | -9.627 | 0.000 | Full Features |
| LASSO | Random Forest | MSE | -9.059 | 0.000 | |
| LASSO | ARIMA | MSE | -8.588 | 0.000 | |
| LASSO | LSTM | MSE | -16.151 | 0.000 | |

| Table 6 - Diebold-Mariano Test Results | | | | | |
|---|---|---|---|---|---|
| Model 1 | Model 2 | Loss Function | DM Statistic | p-value | Feature Set |
| CatBoost | Random Forest | MSE | 0.860 | 0.393 | |
| CatBoost | ARIMA | MSE | -0.906 | 0.368 | |
| CatBoost | LSTM | MSE | -10.510 | 0.000 | |
| Random Forest | ARIMA | MSE | -1.661 | 0.101 | |
| Random Forest | LSTM | MSE | -9.834 | 0.000 | |
| ARIMA | LSTM | MSE | -9.139 | 0.000 | |
| LASSO | CatBoost | MSE | -15.960 | 0.000 | LASSO Selected Features |
| LASSO | Random Forest | MSE | -13.774 | 0.000 | |
| LASSO | ARIMA | MSE | -19.111 | 0.000 | |
| LASSO | LSTM | MSE | 8.874 | 0.000 | |
| CatBoost | Random Forest | MSE | 1.774 | 0.080 | |
| CatBoost | ARIMA | MSE | -8.174 | 0.000 | |
| CatBoost | LSTM | MSE | 23.198 | 0.000 | |
| Random Forest | ARIMA | MSE | -9.626 | 0.000 | |
| Random Forest | LSTM | MSE | 16.582 | 0.000 | |
| ARIMA | LSTM | MSE | 26.206 | 0.000 | |

Note: This table presents Diebold-Mariano (DM) test results based on mean squared prediction errors (MSE). Each row compares the forecast accuracy of two models. The sign of the DM statistic indicates the direction of performance: a negative value means that Model 1 achieved lower MSE than Model 2, while a positive value indicates the opposite. Statistical significance is assessed at the 5% level; p-values below 0.05 indicate a meaningful difference in predictive accuracy.

The results show clear differences in predictive power between the models, especially when considering the impact of variable selection. The version of the LSTM model that used variables selected in advance by LASSO outperformed the others in most comparisons, with statistically significant results. In all cases, for LASSO-Selected Variables, the test statistics were significant at the 5% level,

which means the LSTM had consistently lower mean squared errors compared to the other models, especially when compared to ARIMA, Random Forest, and CatBoost.

This stronger performance of the LSTM-LASSO setup supports what Sezer et al. (2020) found, showing that recurrent neural networks can capture dynamic and nonlinear patterns more effectively when combined with appropriate dimensionality reduction techniques. In this study, the combo worked especially well for carbon market data, which tends to be volatile and influenced by many external shocks.

Interestingly, the linear LASSO model also showed strong results, outperforming all models except LSTM-LASSO. This suggests that a linear model with L1 regularization is still able to capture a good part of the explainable variation in the data, which makes LASSO surprisingly competitive. When both models used the reduced set selected by LASSO, the LSTM regained the lead against LASSO ($DM = 8.874$, $p = 0.000$), highlighting the strength of neural networks when paired with effective dimensionality reduction.

On the other hand, the tree-based models (Random Forest and CatBoost) performed in the middle: better than ARIMA, but not quite as strong as LASSO and LSTM-LASSO.

ARIMA, as expected, was the least effective model overall. It consistently underperformed in almost every comparison, regardless of which variables were used. This confirms the model's limitations in markets with complex structures and strong exposure to external shocks, as already discussed by Han et al. (2019) and Aatola et al. (2013).

Beyond statistical significance, the actual DM test values reinforce the size of the performance gaps. The largest differences were seen between ARIMA and LSTM-LASSO, followed by comparisons between LASSO-LSTM and the ensemble models. The fact that p-values stayed below 0.05 across several windows and setups supports how solid these results are.

In short, the Diebold-Mariano test not only confirms the previous results but makes them stronger by showing that the differences in performance are not due to chance or sample instability. In volatile markets like carbon credits, this kind of evidence is essential to support the use of more advanced forecasting techniques. The combination of neural networks and variable selection does not just deliver

better average results, it does so with statistical significance and consistency, which makes it a highly recommended approach for this type of forecasting problem.

# 5 Conclusions

The results revealed consistent patterns about how well the forecasting models worked in the European carbon market, and they also showed how important the methodological choices were, especially in variable selection and in dealing with complex time dynamics. This section brings together the main findings, based on the goals defined in Section 1 and the academic literature, focusing on what the results mean for forecasting in volatile and regulated markets.

The main empirical takeaway was how much the models improved after applying LASSO for variable selection. The Machine Learning and Neural Network models did not perform greatly when using the full set of variables, often showing signs of overfitting, but once LASSO regularization selection was applied, the accuracy, stability, and simplicity improved a lot. This effect was most visible with the LSTM model, which went from underperforming to becoming the best overall, especially when tested in both fixed-window and expanding-window cross-validation setups, and with DM test.

This is in line with what Sezer et al. (2020) and Smyl (2020) found: deep neural networks can be highly effective in financial time series forecasting, as long as you use proper complexity control techniques like regularization, smart preprocessing, and structured validation. The combination of LASSO and LSTM worked well here to detect hidden patterns in a market shaped by complex interactions between macroeconomic, financial, regulatory, and geopolitical factors.

Tree-based models like CatBoost and Random Forest did not lead the rankings, but they still showed solid and stable performance across both expanding and fixed window setups. This resilience, attributed to ensemble techniques like bagging and boosting (Breiman, 2001; Prokhorenkova et al., 2018), makes them useful alternatives in situations with a lot of noise and multicollinearity. However, their ability to handle multicollinearity may be somewhat limited in multivariate time series settings, where collinearity across lags and shifting regimes poses additional challenges. This is further supported by the fact that both models tended to perform slightly better when using the subset of variables selected by LASSO, suggesting that even ensemble methods benefit from prior dimensionality reduction in complex time series context. Even so, their overall consistency makes them

valuable when interpretability and robustness matter just as much as raw predictive accuracy.

Interestingly, LASSO as a linear model also held up well and, in many cases, performed better than non-linear and non-parametric models like CatBoost and Random Forest, as shown especially in out-of-sample $R^2$ and the Diebold-Mariano test. This suggests that simpler models, when well specified, still have a solid place in more advanced forecasting modelling.

One limitation observed was the performance drop of the LASSO model under the fixed rolling window setup when used with the reduced set of variables. Unlike in the expanding window configuration, where LASSO benefited from accumulating more data over time, its performance worsened here, likely due to the combined effect of smaller training sizes, reduced feature space, and greater exposure to structural breaks. This suggests that linear models with regularization may require both larger samples and broader information sets to maintain stability in dynamic environments with shifting regimes.

ARIMA consistently fell behind. Its univariate, linear structure was not enough to handle the complexity of the carbon market, confirming what Han et al. (2019) and Lin and Zhang (2022) noted about the limitations of traditional models in settings with structural shifts and multiple sources of volatility.

Applying the Diebold-Mariano test added another layer of statistical rigor to the analysis, confirming that the performance gains from LSTM-LASSO were statistically significant, not just random noise or luck. The consistency of these results makes it very relevant for practical use.

Still, it is important to recognize that forecasting in markets with regulatory uncertainty and geopolitical shocks, like the EU ETS, comes with structural limitations. As highlighted by Makridakis et al. (2020) and Smyl (2020), even advanced models can struggle with long-term predictions or sudden regime changes. This study tried to reduce those limitations through rolling window validation, complexity control, and variable selection. But it is clear that the carbon market's volatility remains a major challenge.

To sum up, the results clearly answer a central research question: hybrid models that include dimensionality reduction and robust validation perform better than traditional approaches when forecasting the S&P Carbon Credit EUA Index. It is not about finding the one perfect model, but about understanding how

combining methods and testing multiple approaches helps address the complexity of the problem.

This has direct implications for real-world use: the more stable forecasts of regularized models, the strong performance of LSTM with selected variables, and the reliability of LASSO as both filter and predictor offer practical tools for tasks like hedging, strategy planning, and climate policy evaluation in dynamic regulatory environments.

## 5.1. Practical Implications and Potential Applications

The insights developed in this research extend beyond theory. They carry practical weight, especially for regulators, policymakers, institutional investors, and risk managers who navigate the complexities of regulated carbon markets.

From a policy perspective, the models developed here (especially LASSO and LSTM paired with LASSO) offer a solid tool to anticipate carbon price movements triggered by regulatory or geopolitical shocks. That political and regulatory uncertainty variables were consistently selected is telling. It suggests that these qualitative factors actively shape market expectations. In that sense, public agencies and climate policy institutions could adopt similar forecasting tools to track sentiment, gauge the potential impact of upcoming legislation, or run scenario simulations, like future reforms to the EU ETS or the rollout of CBAM.

As carbon finance continues to expand, with more products like ETFs, carbon-linked derivatives, and sustainability-focused credit instruments entering the market, these results gain relevance in financial strategy. ESG-focused asset managers and large institutional investors, in particular, stand to benefit from improved predictive tools. When regulations shift or uncertainty spikes, better foresight becomes not just useful, but essential.

Incorporating these forecasts into hedging strategies or pricing models allows investors to align risk exposure with expected returns, especially under emerging regulatory frameworks like the European Green Deal or the disclosure principles outlined by the Task Force on Climate-related Financial Disclosures (TCFD). Risk is no longer just about price, it is about policy, reputation, and environmental accountability.

And then there is the private sector. For firms directly subject to carbon market regulations (energy producers, airlines, heavy industries) access to more accurate carbon price forecasts has direct business implications. It can inform investment decisions, cost management strategies, and help quantify future environmental liabilities. In other words, better forecasting helps them remain competitive while meeting their decarbonization goals.

In summary, the models analyzed in this study do more than contribute to academic forecasting literature. They offer practical tools for navigating the regulatory, economic, and political risks surrounding carbon markets.

## 5.2. Path for Future Research

This study adds to the literature by proposing a robust, replicable framework for forecasting carbon prices, one that does not just apply to the EU ETS. By combining variable selection with hybrid modeling techniques and statistical validation, it creates a reference point for both researchers and professionals navigating the complexities of carbon finance.

Looking ahead, there are several promising directions for future work:

(i) One path involves exploring hierarchical Bayesian models or temporal transformers. These approaches may offer better tools for capturing long-range dependencies and accounting for deeper structural changes over time.

(ii) Another is expanding the use of mixed-frequency models, which allow researchers to incorporate indicators released at different temporal resolutions, such as daily energy prices and monthly macroeconomic data, without depending on arbitrary aggregation.

(iii) A third area worth pursuing would be analyzing the impact of climate-related events and policy announcements with higher temporal granularity. Things like EU ETS reform signals or CBAM regulatory milestones could offer new insight if studied with finer time resolution.

## 6 References


AATOLA, Piia; OLLIKAINEN, Markku; TOPPINEN, Anne. Price determination in the EU ETS market: Theory and econometric analysis with market fundamentals. **Energy Economics**, v. 36, p. 380-395, 2013.

ALBEROLA, Emilie; CHEVALLIER, Julien; CHÈZE, Benoît. Price drivers and structural breaks in European carbon prices 2005–2007. **Energy policy**, v. 36, n. 2, p. 787-797, 2008.

BAKER, Malcolm et al. **Financing the response to climate change: The pricing and ownership of US green bonds**. National Bureau of Economic Research, 2018.

BAKER, Scott R.; BLOOM, Nicholas; DAVIS, Steven J. Measuring economic policy uncertainty. **The quarterly journal of economics**, v. 131, n. 4, p. 1593-1636, 2016.

BAKER, Scott R. et al. **Policy news and stock market volatility**. National Bureau of Economic Research, 2019.

BAKER, Scott R. et al. Twitter-derived measures of economic uncertainty. 2021.

BATTISTON, Stefano et al. A climate stress-test of the financial system. **Nature Climate Change**, v. 7, n. 4, p. 283-288, 2017.

BENZ, Eva; TRÜCK, Stefan. Modeling the price dynamics of CO2 emission allowances. **Energy Economics**, v. 31, n. 1, p. 4-15, 2009.

BERGMEIR, Christoph; BENÍTEZ, José M. On the use of cross-validation for time series predictor evaluation. **Information Sciences**, v. 191, p. 192-213, 2012.

BERGMEIR, Christoph; HYNDMAN, Rob J.; KOO, Bonsoo. A note on the validity of cross-validation for evaluating autoregressive time series prediction. **Computational Statistics & Data Analysis**, v. 120, p. 70-83, 2018.

BONTEMPI, Maria Elena et al. EURQ: A new web search-based uncertainty index. **Economica**, v. 88, n. 352, p. 969-1015, 2021.

BOX, GEORGE EP; JENKINS, Gwilym M. Time series analysis: Forecasting and control. **San Francisco: Holden-Day**, 1970.

BREDIN, Don; MUCKLEY, Cal. An emerging equilibrium in the EU emissions trading scheme. **Energy Economics**, v. 33, n. 2, p. 353-362, 2011.

BREIMAN, Leo. Random forests. **Machine learning**, v. 45, p. 5-32, 2001.

BRANGER, Frédéric; QUIRION, Philippe; CHEVALLIER, Julien. Carbon leakage and competitiveness of cement and steel industries under the EU ETS: much ado about nothing. **The Energy Journal**, v. 37, n. 3, 2016.

BROWNLEE, Jason. **Deep learning for time series forecasting: predict the future with MLPs, CNNs and LSTMs in Python**. Machine Learning Mastery, 2018.

BUA, Giovanna et al. Transition versus physical climate risk pricing in European financial markets: A text-based approach. **The European Journal of Finance**, v. 30, n. 17, p. 2076-2110, 2024.

CALEL, Raphael. Carbon markets: a historical overview. **Wiley Interdisciplinary Reviews: Climate Change**, v. 4, n. 2, p. 107-119, 2013.

CALDARA, Dario; IACOVIELLO, Matteo. Measuring geopolitical risk. **American economic review**, v. 112, n. 4, p. 1194-1225, 2022.

CERQUEIRA, Vitor; TORGO, Luis; MOZETIČ, Igor. Evaluating time series forecasting models: An empirical study on performance estimation methods. **Machine Learning**, v. 109, n. 11, p. 1997-2028, 2020.

CHAI, Tianfeng; DRAXLER, Roland R. Root mean square error (RMSE) or mean absolute error (MAE)? Arguments against avoiding RMSE in the literature. **Geoscientific model development**, v. 7, n. 3, p. 1247-1250, 2014.

CHEVALLIER, Julien. A model of carbon price interactions with macroeconomic and energy dynamics. **Energy economics**, v. 33, n. 6, p. 1295-1312, 2011.

CRETI, Anna; JOUVET, Pierre-André; MIGNON, Valérie. Carbon price drivers: Phase I versus Phase II equilibrium?. **Energy Economics**, v. 34, n. 1, p. 327-334, 2012.

DAI, Peng-Fei et al. The impact of economic policy uncertainties on the volatility of European carbon market. **Journal of Commodity Markets**, v. 26, p. 100208, 2022.

DANG, Tam Hoang-Nhat et al. Measuring the energy-related uncertainty index. **Energy Economics**, v. 124, p. 106817, 2023.

DAVIS, Steven J. **An index of global economic policy uncertainty**. National Bureau of Economic Research, 2016.

DIEBOLD, Francis X.; MARIANO, Roberto S. Comparing Predictive Accuracy. **Journal of Business & Economic Statistics**, v. 13, n. 3, p. 253-263, 1995.

ELLERMAN, A. Denny; BUCHNER, Barbara K. The European Union emissions trading scheme: origins, allocation, and early results. 2007.

**EUROPEAN COMMISSION.** *Proposal for a regulation of the European Parliament and of the Council establishing a carbon border adjustment mechanism (COM(2021) 564 final).* Brussels: European Commission, 2021. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0564. Accessed on: May 21, 2025.

FATICA, Serena; PANZICA, Roberto; RANCAN, Michela. The pricing of green bonds: are financial institutions special?. **Journal of Financial Stability**, v. 54, p. 100873, 2021.

GAVRIILIDIS, Konstantinos. Measuring climate policy uncertainty. **Available at SSRN 3847388**, 2021.

GHYSELS, Eric; SANTA-CLARA, Pedro; VALKANOV, Rossen. Predicting volatility: getting the most out of return data sampled at different frequencies. **Journal of Econometrics**, v. 131, n. 1-2, p. 59-95, 2006.

GHANI, Usman et al. Climate change and volatility forecasting: Novel insights from sectoral indices. **Journal of Climate Finance**, v. 6, p. 100034, 2024.

GOODFELLOW, Ian et al. **Deep learning**. Cambridge: MIT press, 2016.

GULEN, Huseyin; ION, Mihai. Policy uncertainty and corporate investment. **The Review of financial studies**, v. 29, n. 3, p. 523-564, 2016.

HAMILTON, James D. A new approach to the economic analysis of nonstationary time series and the business cycle. **Econometrica: Journal of the econometric society**, p. 357-384, 1989.

HAN, Meng et al. Forecasting carbon prices in the Shenzhen market, China: The role of mixed-frequency factors. **Energy**, v. 171, p. 69-76, 2019.

HARRIS, Richard; SOLLIS, Robert. **Applied Time Series Modelling and Forecasting**. 2003.

HARVEY, David; LEYBOURNE, Stephen; NEWBOLD, Paul. Testing the equality of prediction mean squared errors. **International Journal of forecasting**, v. 13, n. 2, p. 281-291, 1997.

HASTIE, Trevor et al. **The elements of statistical learning: data mining, inference, and prediction**. New York: springer, 2009.

HINTERMANN, Beat; PETERSON, Sonja; RICKELS, Wilfried. Price and Market Behavior in Phase II of the EU ETS: A Review of the Literature. **Review of Environmental Economics and Policy**, 2016.

HOCHREITER, Sepp; SCHMIDHUBER, Jürgen. Long short-term memory. **Neural computation**, v. 9, n. 8, p. 1735-1780, 1997.

HOU, Yali; WANG, Qunwei; TAN, Tao. Prediction of carbon dioxide emissions in China using shallow learning with cross validation. **Energies**, v. 15, n. 22, p. 8642, 2022.

HUANG, Yumeng et al. A hybrid model for carbon price forecasting using GARCH and long short-term memory network. **Applied Energy**, v. 285, p. 116485, 2021.

HYNDMAN, Rob J.; ATHANASOPOULOS, George. **Forecasting: principles and practice**. OTexts, 2018.

HYNDMAN, Rob J.; KOEHLER, Anne B. Another look at measures of forecast accuracy. **International journal of forecasting**, v. 22, n. 4, p. 679-688, 2006.

JENKO, Jakob; COSTA, Joao Pita. Using temporal fusion transformer predictions to maximise use of renewable energy sources. In: **2024 International Workshop on Artificial Intelligence and Machine Learning for Energy Transformation (AIE)**. IEEE, 2024. p. 1-10.

JI, Lei et al. Carbon futures price forecasting based with ARIMA-CNN-LSTM model. **Procedia Computer Science**, v. 162, p. 33-38, 2019.

JIN, Jiayu et al. The hedging effect of green bonds on carbon market risk. **International Review of Financial Analysis**, v. 71, p. 101509, 2020.

KWIATKOWSKI, Denis et al. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?. **Journal of econometrics**, v. 54, n. 1-3, p. 159-178, 1992.

KOCH, Nicolas et al. Causes of the EU ETS price drop: Recession, CDM, renewable policies or a bit of everything?—New evidence. **Energy Policy**, v. 73, p. 676-685, 2014.

KUHN, Max et al. **Applied predictive modeling**. New York: Springer, 2013.

LI, Peng et al. Time-varying impacts of carbon price drivers in the EU ETS: a TVP-VAR analysis. **Frontiers in Environmental Science**, v. 9, p. 651791, 2021.

LIM, Bryan et al. Temporal fusion transformers for interpretable multi-horizon time series forecasting. **International Journal of Forecasting**, v. 37, n. 4, p. 1748-1764, 2021.

LIN, Boqiang; ZHANG, Chongchong. Forecasting carbon price in the European carbon market: The role of structural changes. **Process Safety and Environmental Protection**, v. 166, p. 341-354, 2022.

LIU, Liping; LÜ, Zheng. Policy uncertainty, geopolitical risks and China's carbon neutralization. **Carbon Management**, v. 14, n. 1, p. 2251929, 2023.

LIU, Sha et al. Fluctuations and Forecasting of Carbon Price Based on A Hybrid Ensemble Learning GARCH-LSTM-Based Approach: A Case of Five Carbon Trading Markets in China. **Sustainability**, v. 16, n. 4, p. 1588, 2024.

MAKRIDAKIS, Spyros; SPILIOTIS, Evangelos; ASSIMAKOPOULOS, Vassilios. Statistical and Machine Learning forecasting methods: Concerns and ways forward. **PloS one**, v. 13, n. 3, p. e0194889, 2018.

MAKRIDAKIS, Spyros; SPILIOTIS, Evangelos; ASSIMAKOPOULOS, Vassilios. The M4 Competition: 100,000 time series and 61 forecasting methods. **International Journal of Forecasting**, v. 36, n. 1, p. 54-74, 2020.

MAKRIDAKIS, Spyros; SPILIOTIS, Evangelos; ASSIMAKOPOULOS, Vassilios. M5 accuracy competition: Results, findings, and conclusions. **International Journal of Forecasting**, v. 38, n. 4, p. 1346-1364, 2022.

NARASSIMHAN, Easwaran et al. Carbon pricing in practice: A review of existing emissions trading systems. **Climate Policy**, v. 18, n. 8, p. 967-991, 2018.

NIU, Huawei; LIU, Tianyu. Forecasting the volatility of European Union allowance futures with macroeconomic variables using the GJR-GARCH-MIDAS model. **Empirical Economics**, v. 67, n. 1, p. 75-96, 2024.

OBERNDORFER, Ulrich. EU emission allowances and the stock market: evidence from the electricity industry. **Ecological Economics**, v. 68, n. 4, p. 1116-1126, 2009.

ORESHKIN, Boris N. et al. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. **arXiv preprint arXiv:1905.10437**, 2019.

PASZKE, A. Pytorch: An imperative style, high-performance deep learning library. **arXiv preprint arXiv:1912.01703**, 2019.

PASTOR, Lubos; VERONESI, Pietro. Uncertainty about government policy and stock prices. **The journal of Finance**, v. 67, n. 4, p. 1219-1264, 2012.

PEDREGOSA, Fabian et al. Scikit-learn: Machine learning in Python. **the Journal of machine Learning research**, v. 12, p. 2825-2830, 2011.

PERINO, Grischa. New EU ETS Phase 4 rules temporarily puncture waterbed. **Nature Climate Change**, v. 8, n. 4, p. 262-264, 2018.

PERINO, Grischa; WILLNER, Maximilian. Procrastinating reform: The impact of the market stability reserve on the EU ETS. **Journal of Environmental Economics and Management**, v. 80, p. 37-52, 2016.

PESARAN, M. Hashem; TIMMERMANN, Allan. Predictability of stock returns: Robustness and economic significance. **The Journal of Finance**, v. 50, n. 4, p. 1201-1228, 1995.

PROKHORENKOVA, Liudmila et al. CatBoost: unbiased boosting with categorical features. **Advances in neural information processing systems**, v. 31, 2018.

ROGERS, John H.; SUN, Bo; SUN, Tony. US-China tension. **Available at SSRN 4815838**, 2024.

SANTIKARN, Marissa et al. State and trends of carbon pricing 2021. 2021.

SEZER, Omer Berat; GUDELEK, Mehmet Ugur; OZBAYOGLU, Ahmet Murat. Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. **Applied soft computing**, v. 90, p. 106181, 2020.

SHMUELI, Galit. To explain or to predict?. 2010.

SMYL, Slawek. A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. **International journal of forecasting**, v. 36, n. 1, p. 75-85, 2020.

TASHMAN, Leonard J. Out-of-sample tests of forecasting accuracy: an analysis and review. **International journal of forecasting**, v. 16, n. 4, p. 437-450, 2000.

TIBSHIRANI, Robert. Regression shrinkage and selection via the lasso. **Journal of the Royal Statistical Society Series B: Statistical Methodology**, v. 58, n. 1, p. 267-288, 1996.

VERDE, Stefano F. et al. Free allocation rules in the EU emissions trading system: what does the empirical literature show?. **Climate policy**, v. 19, n. 4, p. 439-452, 2019.

WANG, Xiaoqing et al. Volatility in Carbon Futures Amid Uncertainties: Considering Geopolitical and Economic Policy Factors. **Journal of Futures Markets**, 2025.

WU, Han; DU, Pei. Dual-stream transformer-attention fusion network for short-term carbon price prediction. **Energy**, v. 311, p. 133374, 2024.

XU, Hua et al. Carbon price forecasting with complex network and extreme learning machine. **Physica A: Statistical Mechanics and its Applications**, v. 545, p. 122830, 2020.

ZHANG, G. Peter. Time series forecasting using a hybrid ARIMA and neural network model. **Neurocomputing**, v. 50, p. 159-175, 2003.

ZHANG, Zuocheng et al. Green finance and carbon emission reduction: A bibliometric analysis and systematic review. **Frontiers in Environmental Science**, v. 10, p. 929250, 2022.

ZHAO, Xin et al. Usefulness of economic and energy data at different frequencies for carbon price forecasting in the EU ETS. **Applied Energy**, v. 216, p. 132-141, 2018.

ZHU, Bangzhu et al. Carbon price forecasting with a hybrid Arima and least squares support vector machines methodology. **Pricing and forecasting carbon markets: Models and empirical analyses**, p. 87-107, 2017.

ZOU, Hui; HASTIE, Trevor. Regularization and variable selection via the elastic net. **Journal of the Royal Statistical Society Series B: Statistical Methodology**, v. 67, n. 2, p. 301-320, 2005.