RESEARCH AND REVIEWS IN ECONOMICS AND MANAGEMENT
Research Article
Volume 2 | issue 5

The Price of Greenwashing: Algorithmic Verification and Market Discipline using Conformal Machine Learning

Sourav Bose[1*] & Taoufik Bouraoui[2]

[1]*Head Of AI, Unsw Dba Candidate, Rennes School Of Business 2 Rue Robert D'arbrissel, Cs 76522 35065, Rennes Cedex, France.*
[2]*Professor of Finance, Rennes School Of Business 2 Rue Robert D'arbrissel, Cs 76522 35065, Rennes Cedex, France*

[*]**Corresponding Author:** Sourav Bose, Head Of AI, Unsw Dba Candidate, Rennes School Of Business 2 Rue Robert D'arbrissel, Cs 76522 35065, Rennes Cedex, France.

Submitted: 27 August 2026 **Accepted:** 03 September 2026 **Published:** 10 September 2026

Citation: Bose, S., Bouraoui, T. (2026).The Price of Greenwashing: Algorithmic Verification and Market Discipline using Conformal Machine Learning. J of Research and Reviews in Economics and Management. 2(5), 01-23.

Abstract
While corporate sustainability mandates are expanding, the systemic reliance on self-reported emissions data exposes financial markets to pervasive greenwashing. Current literature predominantly utilizes subjective commercial ESG ratings or textual sentiment analysis to detect this deception, leaving a critical econometric gap in objectively quantifying physical climate realities. To resolve this information asymmetry, we engineer a novel algorithmic verification pipeline. Fusing U.S. SEC financial fundamentals with facility-level EPA greenhouse gas registries, we train a hyperparameter-optimized Extreme Gradient Boosting (XGBoost) architecture to predict physical baseline emissions. To rigorously quantify epistemic uncertainty and account for extreme asset-class heteroskedasticity, we apply a Mondrian Conformal Prediction framework, generating mathematically guaranteed, firm-specific safety margins. The shortfall between self-reported emissions and this conformal baseline is standardized into a novel Conformal-Weighted Continuous Divergence (CWCD) metric. Utilizing a cross-sectional lead-lag econometric design with high-dimensional industry fixed effects, we evaluate the directional, leading predictive association of algorithmic divergence on subsequent market valuation (Tobin's Q) and operational profitability (ROA). The empirical estimations reveal a robust predictive association indicative of market discipline: algorithmic emissions divergence exhibits a severe, statistically significant negative relationship with subsequent forward valuation and operational profitability. Providing robust evidence against the market blindness hypothesis, this study strongly suggests that institutional capital possesses the mechanisms to actively price environmental deception as a leading indicator of fundamental corporate mismanagement. These findings provide the quantitative justification necessary for regulatory authorities and asset managers to deploy algorithmic auditing infrastructure at scale.

Keywords: Corporate Greenwashing, Climate Finance, Conformal Prediction, Market Discipline, Empirical Asset Pricing, Machine Learning in Finance, ESG Disclosures

JEL Classifications: G14, G32, Q56, C55, C58

Introduction

The global transition toward sustainable capital markets is fundamentally constrained by the persistent information asymmetry inherent in corporate greenwashing. As sustainability mandates expand, corporate entities frequently exploit voluntary, self-audited Environmental, Social, and Governance (ESG) reporting frameworks to strategically misrepresent their physical climate realities. Existing financial literature predominantly attempts to document this deception through subjective commercial ESG rating agencies or textual sentiment analysis. For instance, foundational studies such as establish that markets price self-reported carbon risks, while expose the profound measurement divergence and unreliability inherent in using commercial ESG ratings to proxy these risks. However, despite these seminal efforts, a critical econometric gap remains: while the literature proves the market prices reported emissions and subjective ESG scores, how can financial markets accurately price corporate climate risk when investors lack an objective, verifiable baseline of actual physical operational emissions? [1, 2].

This paper exploits advanced non-linear machine learning to investigate the systemic pricing of physical climate deception within US equity markets. Unlike traditional approaches that rely on backward-looking commercial ratings or easily manipulated public relations narratives, this study explicitly leverages the deterministic constraints of corporate financial fundamentals. By mathematically linking a firm's operational scale, capital expenditure, and profitability to its physical emissions footprint, we establish an objective algorithmic baseline of climate reality. We deploy this baseline to resolve a critical theoretical tension: whether extreme information asymmetry renders environmental deception an unpriced market externality, or whether institutional capital possesses the mechanisms to actively discipline strategic decoupling.

To quantify this deception, we architect a multi-tiered machine learning and econometric pipeline to evaluate a globally fused panel of facility-level EPA greenhouse gas registries and SEC financial disclosures. At the algorithmic level, we utilize a hyperparameter-optimized Extreme Gradient Boosting (XGBoost) architecture to predict theoretical baseline emissions. Crucially, because corporate emissions exhibit extreme right-skewed Gamma distributions and pronounced asset-class heterogeneity, traditional point predictions structurally fail. Therefore, we deploy a Mondrian Conformal Prediction framework to generate mathematically guaranteed, firm-specific uncertainty margins. At the econometric level, we standardize the shortfall between reported emissions and this conformal safety net into a novel Conformal-Weighted Continuous Divergence (CWCD) score. To isolate temporal causality, we evaluate this metric using a cross-sectional lead-lag Ordinary Least Squares (OLS) regression, leveraging high-dimensional Industry Fixed Effects to absorb unobservable, sector-specific valuation imbalances.

Our baseline temporal estimates reveal a robust, highly significant market disciplinary mechanism. After rigorously controlling for baseline financial fundamentals and isolating intra-industry dynamics, the algorithmic divergence penalty demonstrates a severe, negative temporal drag on both subsequent market valuation (Tobin's Q, $p < 0.01$) and operational profitability (ROA, $p < 0.01$). Rather than indicating a market failure, we diagnose this as empirical proof of formidable market efficiency. The US equities market aggressively penalizes algorithmic emissions divergence, revealing that while corporate greenwashing may evade regulatory scrutiny, it cannot evade the pricing mechanism of institutional capital.

This study makes three primary contributions to the intersecting literatures of empirical corporate finance, applied machine learning, and environmental economics:

First, we introduce a novel, objective proxy for environmental deception. By engineering the CWCD metric, we empirically validate a methodology that standardizes emissions shortfalls against Mondrian marginal bounds, providing a superior, fundamentals-driven alternative to subjective commercial ESG ratings.

Second, we provide robust empirical evidence supporting the efficacy of market discipline regarding physical climate risk. We demonstrate that despite severe information asymmetry and the lack of algorithmic auditing infrastructure, institutional investors are successfully utilizing alternative heuristics to proxy operational deception. The massive valuation penalty levied against divergent firms proves that the market actively prices the fundamental mismanagement inherent in strategic environmental decoupling.

Third, we bridge the gap between machine learning and causal asset pricing. By demonstrating that this conformal machine learning architecture effectively proxies the localized climate risk parameters enforced by the market, this research lays the structural empirical groundwork necessary for institutional investors and regulatory authorities (such as the SEC and ISSB) to deploy algorithmic verification at scale.

Ultimately, this study demonstrates that localized, performative sustainability rhetoric is systematically neutralized by the broader financial ecosystem. By shifting the verification burden to algorithmic machine learning frameworks, we establish the empirical truth required to accurately map and price systemic macroeconomic climate risks.

The remainder of this paper is organized as follows. Section 2 reviews the existing literature on information asymmetry and develops the theoretical hypotheses. Section 3 details the data engineering, the XGBoost and Mondrian Conformal Prediction architecture, and the causal lead-lag econometric specification. Section 4 presents the empirical results, including the conformal uncertainty bounds, the explainable AI (XAI) and robustness diagnostics, and the final causal market valuation models. Section 5 discusses the theoretical and policy implications of algorithmic verification for institutional capital and regulatory bodies. Section 6 concludes the paper.

Literature Review & Hypotheses Development

Information Asymmetry and the Economics of Greenwashing

The proliferation of corporate sustainability mandates has catalysed a corresponding surge in "greenwashing"—the strategic misrepresentation of environmental performance. attribute this phenomenon to a complex interplay of endogenous organizational drivers and an exogenous, fundamentally lax regulatory environment, noting that such deception critically undermines investor confidence. In the absence of stringent, algorithmically verifiable disclosure mandates, firms exploit information asymmetry to engage in strategic decoupling, wherein external ESG rhetoric diverges sharply from physical operational reality[3,4]. Furthermore, while the threat of activist or regulatory auditing theoretically deters overt falsification, it frequently induces firms to adopt selective disclosure strategies [5]. Formalize this behaviour along a spectrum ranging from "greenwashing" to "brownwashing," a dynamic heavily dictated by the prevailing regulatory architecture [6].

Comprehensively analyse the economic implications of mandatory CSR reporting, emphasizing that while capital markets react to sustainability disclosures, the enforcement and standardization of these metrics remain structurally deficient [7]. The difficulty of enforcing environmental standards is further highlighted by, who demonstrate that while environmental violations incur market value losses, these are almost entirely driven by direct legal penalties rather than reputational damages. Consequently, systemic reliance on self-reported, voluntary ESG disclosures creates a critical vulnerability in global capital markets, enabling environmental deception to propagate without material financial consequence [8].

The Measurement Void: ESG Rating Divergence and Climate Risk Pricing

To navigate this pervasive information asymmetry, institutional capital has historically outsourced verification to commercial ESG rating agencies. However, this reliance introduces a profound measurement gap. Formally document this in Aggregate Confusion: The Divergence of ESG Ratings, demonstrating that ratings across prominent agencies exhibit average pairwise correlations ranging from merely 38% to 71%. Their decomposition reveals that measurement divergence accounts for 56% of this discrepancy, rendering commercial ESG scores inherently subjective and unreliable as proxies for physical climate realities. This lack of convergent validity is corroborated by who warn that low commensurability among raters severely impairs strategic asset allocation. Moreover, find that ESG rating disagreement correlates with higher stock returns, suggesting the existence of a risk premium driven largely by the environmental dimension [2-10].

Parallel to the ESG ratings literature, empirical asset pricing scholars have begun investigating the market's pricing of climate risk [11]. Demonstrate that investors demand a carbon risk premium, finding that higher total CO2 emissions correlate with higher cross-sectional stock returns. Similarly, establish a negative firm-value effect for carbon emissions, while theorize that in equilibrium, green assets command lower expected returns due to investor preferences and climate hedging. and further highlight how institutional investors actively attempt to hedge against climate-related shocks. However, as critiques, the over-reliance on aggregated ESG scores often obscures fundamental long-term value drivers. A critical econometric blind spot thus remains: while the market may price reported carbon emissions, it lacks the objective infrastructure to price emissions deception—the divergence between reported figures and a firm's true physical baseline [12-17].

Bridging the Methodological Gap: Machine Learning and Conformal Prediction

To resolve the limitations of subjective ESG ratings and simple linear models, recent literature has increasingly turned to advanced computational architectures. and established the foundational supremacy of machine learning in empirical asset pricing and causal econometrics, demonstrating that tree-based architectures capture complex, non-linear predictor interactions that traditional frameworks structurally miss. However, the vanguard of the field has rapidly evolved beyond basic point predictions. As comprehensively surveyed by, the modern transition from traditional econometrics to machine learning in asset pricing strictly demands a unified framework that balances high-dimensional predictive power with deep economic interpretability. Concurrently, recent empirical applications, such as optimized machine learning framework, have proven that non-linear ensemble architectures are highly effective at isolating specific corporate features driving ESG greenwashing behaviour [18-21].

Crucially, the efficacy of any algorithmic auditing mechanism depends entirely on the theoretical validity of its predictor variables. Standard environmental economics and neoclassical production theory establish that physical emissions are an unavoidable byproduct of fixed capital formation and operational scale [22]. Financial fundamentals—specifically capital expenditure (CapEx), total asset scale, and return on assets (ROA) act as deterministic constraints on a firm's physical output. As established by theoretical frameworks linking corporate investment to environmental externalities, a firm cannot drastically alter its physical emissions trajectory without a corresponding shift in its capital allocation and asset utilization. Therefore, utilizing a machine learning architecture trained on baseline financial fundamentals provides an objective, mathematically bounded proxy for physical climate reality, effectively neutralizing the rhetorical manipulation inherent in textual sustainability reports [23, 24].

However, standard machine learning architectures typically yield point estimates that lack rigorous uncertainty quantification—a fatal flaw when attempting to price risk in highly volatile equity markets. To establish a mathematically guaranteed safety net around these predictions, the Conformal Prediction framework has emerged as a vital methodological bridge. As pioneered comprehensively by and expanded in their updated synthesis, conformal predictors evaluate the reliability of their own forecasts in a way that is rigorously valid under the assumption of data exchangeability, requiring no strict parametric assumptions regarding the underlying data-generating distribution [25, 26].

While classical conformal prediction merely outputs static prediction sets, recent advancements provide the exact mathematical infrastructure required for advanced financial auditing. Introduced Conformal Predictive Distributions, successfully shifting the paradigm from static bounding intervals to dynamic, non-parametric probability distributions. This capacity to output full predictive distributions with mathematically guaranteed coverage is further optimized for high-dimensional efficiency by Concurrently, the development of Conformal Calibrators allows for the seamless calibration of machine learning outputs into precise probabilistic risk measures [27-29].

Furthermore, empirical financial cross-sections inherently suffer from structural heterogeneity and macroeconomic drift. As demonstrated by, modern Conformal Predictive Systems (CPS) can now dynamically adjust to severe covariate shifts, allowing for mathematically valid inference even when traditional independent and identically distributed (I.I.D.) assumptions fail. By integrating a Mondrian Conformal Prediction architecture to account for this severe heteroskedasticity, this study leverages these modern predictive distributions. This methodology translates theoretical machine learning point predictions not merely into rigid safety margins, but into a calibrated, continuous baseline of corporate climate reality—the necessary econometric foundation for extracting the continuous algorithmic divergence penalty (CWCD) priced by institutional capital [30].

Hypothesis Development

Drawing upon these intersecting literatures, this study tests the efficacy of market discipline in pricing environmental deception. The literature presents two competing theoretical frameworks regarding the market's capacity to penalize algorithmic emissions divergence.

The Efficient Market and Signalling framework posit that, despite the lack of standardized reporting, institutional investors possess the analytical sophistication to proxy corporate climate risks and actively penalize deceptive disclosures [31]. Under this framework, firms exhibiting high algorithmic divergence—indicating a severe shortfall between reported emissions and their financial-fundamental baseline—should suffer a risk premium penalty, manifesting as depressed valuation and constrained operational efficiency.

Hypothesis 1 (Market Discipline): Algorithmic emissions divergence exerts a statistically significant negative drag on subsequent market valuation (Tobin's Q) and operational profitability (ROA). Conversely, the Information Asymmetry and Verification Cost framework argues that without algorithmic auditing infrastructure, the cost for investors to cross-validate self-reported emissions against physical constraints is prohibitively high. If commercial ESG ratings fail to capture this decoupling , investors remain structurally blind to the deception. Under this framework, greenwashing functions as a cost-free, unpriced externality [2].

Hypothesis 2 (Market Blindness): Due to extreme information asymmetry, algorithmic emissions divergence exhibits no causal impact on subsequent market valuation (Tobin's Q) or operational profitability (ROA).

By utilizing the CWCD metric within a cross-sectional lead-lag econometric design, this study empirically adjudicates between these competing frameworks. We test whether the US equities market currently possesses the mechanism to discipline environmental deception, positing that if institutional investors are deploying advanced alternative data to pierce corporate veils, the algorithmic divergence penalty will align precisely with the Market Discipline framework.

Data and Methodology

To overcome the inherent limitations of subjective ESG ratings and investigate the systemic pricing of environmental deception, this study develops a two-stage hybrid methodology. Stage 1 utilizes a machine learning pipeline infused with Mondrian Conformal Prediction to establish an objective, fundamentals-driven baseline of physical climate reality. Stage 2 imports this algorithmic baseline into a cross-sectional lead-lag econometric framework to test for causal market effects. Figure 1 shows the structural progression of this methodology across three functional modules: the initial semantic data engineering and fusion (Box 1), the machine learning tournament and generation of the algorithmic safety net (Box 2), and the final calculation of the CWCD metric for causal inference via lead-lag OLS regression

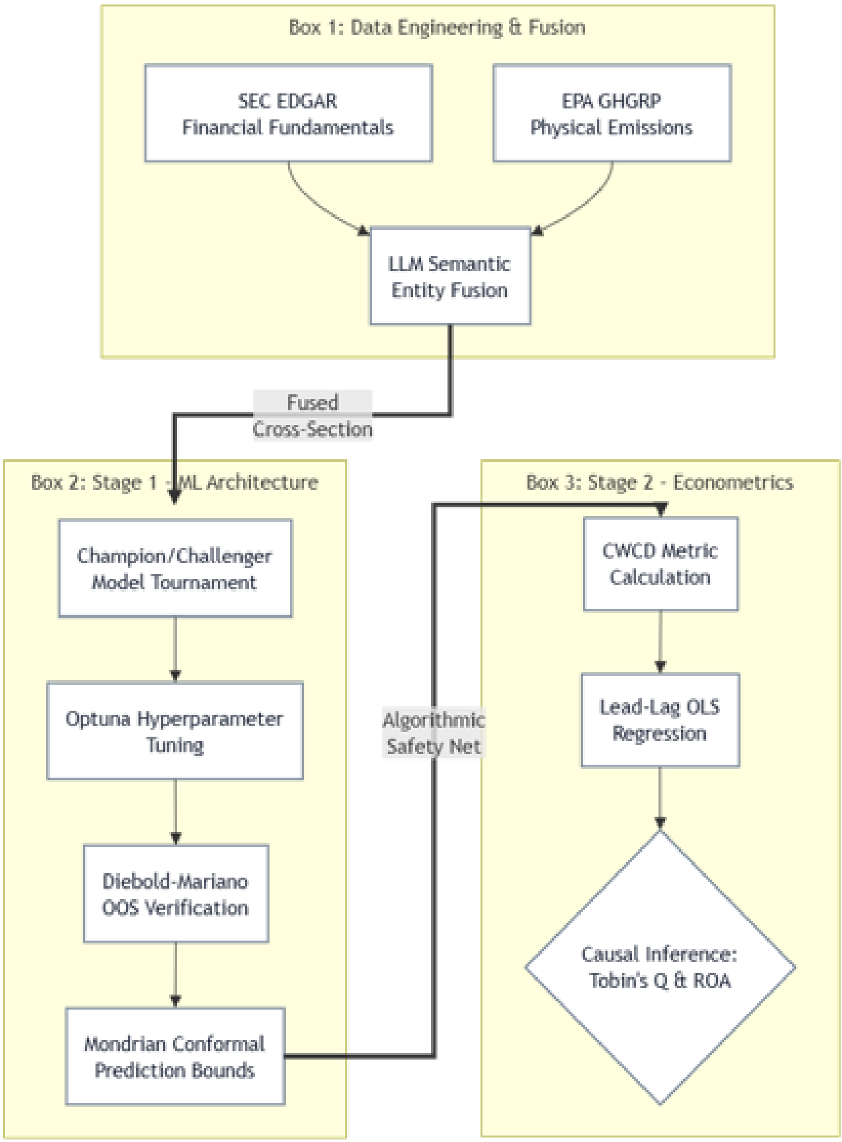


Figure 1: The Two-Stage Algorithmic and Econometric Pipeline

Note: This flowchart shows the overall approach and methodology of the 2-Stage Algorithmic and Econometric pipeline used in this paper

Data Acquisition and Semantic Data Cleaning

The empirical foundation of this study relies on the fusion of highly granular datasets. Because physical carbon emissions are reported at the facility level while financial data is aggregated at the corporate entity level, traditional relational database merges structurally fail due to naming inconsistencies and subsidiary obfuscation. To resolve this, we engineer a multi-layered extraction and semantic fusion pipeline:

Physical Emissions (EPA GHGRP): We automate the extraction of Facility Dimension and Emissions Fact tables from the Environmental Protection Agency's Greenhouse Gas Reporting Program (GHGRP) across the 2020–2023 reporting years. Facility-level emissions are aggregated to the ultimate corporate parent level.

Corporate Identifiers (SEC EDGAR): We query the SEC EDGAR API to retrieve the complete mapping of corporate tickers, CIK identifiers, and official registrant names.

Cascading Entity Resolution Pipeline: To bridge the gap between EPA parent names and SEC registrant names, we deploy a micro-batched Large Language Model (LLM) architecture.

Stage 1 (Syntactic Check): We utilize token-set ratio algorithms to identify direct matches (score ≥90) and reject negligible overlaps (score < 30).

Stage 2 (Semantic Evaluation): Ambiguous "Gray Zone" pairs are routed into a Gemini-3.1-Flash-Lite LLM in micro-batches of 15. The LLM explicitly determines if the pairs represent the exact economic entity, a parent/subsidiary relation, or a corporate rebranding, returning a structured JSON response and a confidence score. Only pairs with a confidence score ≥0.85 are accepted as validated matches.

Feature Engineering and the Lead-Lag Crosswalk

Upon establishing the crosswalk matrix, we engineer a longitudinal panel by performing programmatic extraction of trailing financial data via the Yahoo Finance API, utilizing user-agent mimicking to ensure request stability. To prevent data leakage and isolate the primary trading vehicle, we enforce an empirical deduplication hierarchy that penalizes derivative tickers (e.g., preferred stock, warrants) and unusually long tickers. The subsidiary-level records are then consolidated into absolute Corporate Parent-Year observations.

For each Anchor Year (T), we extract the antecedent (T-1) financial constraints:

Firm Size: Calculated as the natural logarithm of Total Assets.
ROA (Return on Assets): Net Income scaled by Total Assets.
CapEx Intensity: Absolute Capital Expenditures scaled by Total Assets.
Leverage: Total Liabilities scaled by Total Assets.
Market Valuation (Tobin's Q): Calculated utilizing the closing market capitalization within a 10-day window around the fiscal reporting date.

Due to the structural constraints of open-source financial APIs, we pivot the panel into a strictly cross-sectional Lead-Lag Temporal Design (Table 1) focused on the 2023 snapshot. This architecture optimally isolates temporal causality and neutralizes reverse-causality critiques.

Table 1: The Lead-Lag Temporal Architecture

Temporal Phase	Data Source	Variables Extracted	Role in Methodology
Year T-1 (Antecedent)	SEC EDGAR	Firm Size, CapEx, Leverage, ROA, Tobin’s Q	ML Predictor Features (Stage 1)
Year T (Anchor)	EPA GHGRP	Scope 1 Emissions	The Physical Ground Truth (Stage 1)
Year T (Anchor)	Algorithmic	CWCD Score	The Deception Penalty (Stage 2 IV)
Year T+1 (Lead)	Equity Markets	Tobin's Q (leading), ROA (leading)	The Causal Target (Stage 2 DV)

Stage 1: Algorithmic Verification via Conformal Machine Learning

Algorithmic Architecture and OOS Verification

Exploratory Data Analysis (EDA), specifically leveraging LOWESS versus OLS scatter plots, confirms a severe non-linear relationship between baseline constraints and physical emissions. Furthermore, corporate emissions data exhibits extreme right-skewed non-normality, most closely mirroring a Gamma distribution. Because traditional Ordinary Least Squares (OLS) estimation fails under these conditions, deploying an advanced non-linear architecture is required.

To establish the objective physical baseline, we executed a competitive "Champion versus Challenger" cross-sectional tournament utilizing 80% of the dataset for in-sample training. Five candidate architectures (a baseline Gamma Generalized Linear Model (GLM), SVR Log-Link, Random Forest Log-Link, XGBoost Gamma, and LightGBM Gamma) were evaluated. Each candidate utilized Bayesian hyperparameter optimization (Optuna) over 100 trials and 5-Fold cross-validation to minimize the Gamma deviance loss function and Mean Absolute Error (MAE).

Let Y_i represent the reported emissions for firm i, and (Y_i)ˆ represent the predicted emissions. The models learn the feature mapping to minimize the deviance:

$$D(Y,\hat{Y}) = 2\sum_{i=1}^{N}\left(-\ln\left(\frac{Y_i}{\hat{Y}_i}\right)+\frac{Y_i-\hat{Y}_i}{\hat{Y}_i}\right)$$

As detailed in Table 2, the non-linear machine learning architectures vastly outperformed the baseline Gamma Generalized Linear Model (GLM). While the Support Vector Regressor (SVR-LogLink) achieved the lowest cross-validation MAE (14.0566 MT), the tree-based architectures (XGBoost and Random Forest) followed closely, demonstrating exceptional stability in minimizing the Root Mean Squared Error (RMSE).

Table 2: The Gamma-Optimized Cross-Validation Tournament Leaderboard

Model Architecture	CV MAE (Megatonnes)	CV RMSE (Megatonnes)
SVR_LogLink	14.0566	41.1228
XGBoost_Gamma	14.4751	40.9413
RandomForest_LogLink	14.531	41.0996
LightGBM_Gamma	15.1432	41.0305

Gamma_GLM (Baseline)	18.5126	43.1581

Note: Results represent the optimal hyperparameter configurations achieved over 100 Optuna trials utilizing 5-Fold cross-validation on the in-sample dataset.

To objectively adjudicate true predictive superiority, the top models were frozen and evaluated on the intact 20% out-of-sample (OOS) holdout. We utilized a three-way Diebold-Mariano (DM) test to evaluate the variance of the forecast errors between the competing architectures.

Table 3: Out-of-Sample Diebold-Mariano (DM) Predictive Accuracy Tests

Model Comparison	DM Statistic	P-Value	Conclusion
SVR-LogLink vs. Gamma GLM	5.8317	< 0.001	SVR Superior ***
XGBoost vs. Gamma GLM	6.0737	< 0.001	XGBoost Superior ***
SVR-LogLink vs. XGBoost	-1.0156	0.3098	Statistical Tie (n.s.)

Note: The DM test evaluates the null hypothesis of equal predictive accuracy. Significance levels: *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$, n.s. = not significant.

As presented in Table 3, both the SVR and the XGBoost architecture demonstrated a statistically significant reduction in error compared to the baseline ($p < 0.001$). However, when evaluated against each other, the DM test yielded a statistical tie ($p = 0.3098$), indicating indistinguishable predictive variance.

Table 4: Final out-of-Sample Evaluation across Champion, Challenger & Baseline models

Metric	SVR-Loglink	XGBoost	Gamma GLM (baseline)
Mean Absolute Error (OOS)	14.606 MT	13.393 MT	20.806 MT

Because algorithmic auditing requires optimization across multiple dimensions, the final model selection defaulted to absolute accuracy, structural precision, and interpretability:

Absolute Accuracy: As detailed in Table 4, XGBoost achieved the absolute minimum OOS error on the holdout (13.393 MT).

Conformal Efficiency: Given the severe right-skew of the target variable, XGBoost demonstrated superior precision at the extreme tails of the distribution, generating tighter Mondrian marginal bounds for Large-cap entities (±270.98 MT vs. ±278.37 MT).

Mathematical Transparency: Unlike SVRs, which require computationally approximated feature attributions, the XGBoost architecture natively supports mathematically exact Shapley additive explanations (TreeExplainer).

Given its superior absolute out-of-sample accuracy, tail-end precision, and algorithmic transparency, XGBoost was serialized as the core predictive engine for the subsequent econometric specifications.

The Safety Net: Cross-Conformal Mondrian Calibration

A point prediction (($\hat{Y}_i$)) alone is insufficient for corporate auditing due to inherent asset-class heterogeneity. To establish rigorous, finite-sample statistical bounding, we embed the XGBoost predictions within a Mondrian Cross-Conformal Prediction (CV+) framework, leveraging its robust theoretical guarantees under standard exchangeability assumptions, calculating unbiased Out-of-Fold (OOF) residuals across the training splits.

Unlike standard conformal prediction, the Mondrian approach stratifies the data to account for heteroskedasticity. We established static economic thresholds based on Firm Size (log parameters equating to \$3.5 Billion and \$40 Billion in Total Assets) to categorize firms into Small, Medium, and Large strata (S_i). A non-conformity score margin (M_i) is calculated for each bracket to guarantee conditional coverage at $\alpha = 0.05$ (95% confidence). The algorithm outputs a lower bound (L_i) and an upper bound (U_i) such that the true emissions value Y_i falls within the safety margin with 1-α probability:

$$P(Y_i \in [L_i, U_i] \mid S_i) \geq 1 - \alpha$$

Stage 2: Causal Inference and Econometric Specification

The Conformal-Weighted Continuous Divergence (CWCD) Metric

With the physical baseline and conformal boundaries established, we standardize the deception penalty into the novel Conformal-Weighted Continuous Divergence (CWCD) metric.

The CWCD mathematically penalizes a firm only when its self-reported emissions ($Y_{i,t}$) fall suspiciously below the algorithm's predicted baseline (($\hat{Y}_{i,t}$). The magnitude of this penalty is standardized by the firm's specific Mondrian Margin ($M_{i,t}$), ensuring equitable assessment across volatile and stable asset classes. Utilizing a maximum function to truncate over-reporting at zero, the CWCD for firm i at Anchor Year t is formalized as:

$$CWCD_{i,t} = \max\left(0, \frac{\widehat{Y_{i,t}} - Y_{i,t}}{M_{i,t}}\right)$$

Causal Lead-Lag Regression (OLS)

The extracted CWCD feature is imported into a robust causal econometric framework. To adjudicate between the competing theories of Market Discipline and Market Blindness, we test the causal impact of the algorithmic divergence penalty (Year t) on subsequent market valuation (Tobin's Q) and operational profitability (ROA) in Year t+1.

To isolate the unique effect of deception, we explicitly control for the firm's baseline financial fundamentals (X_(i,t)) and deploy high-dimensional Industry Fixed Effects (δ_j) to absorb unobservable, sector-specific valuation imbalances. Furthermore, independent and dependent financial variables are Winsorized at the 1% and 99% levels to neutralize extreme outliers.

The baseline valuation model is specified as:

$$Q_{i,t+1} = \beta_0 + \beta_1 CWCD_{i,t} + \gamma' X_{i,t} + \delta_j + \epsilon_{i,t+1}$$

The secondary autoregressive profitability model is specified as:

$$ROA_{i,t+1} = \beta_0 + \beta_1 CWCD_{i,t} + \gamma' X_{i,t} + \delta_j + \epsilon_{i,t+1}$$

where:

- β_1 is the coefficient of interest, representing the market's pricing of environmental deception.
- $X_{i,t}$ is a vector of controls including Firm Size, CapEx Intensity, Leverage, and baseline ROA.
- δ_j represents categorical Industry Fixed Effects to prevent omitted variable bias.
- $\epsilon_{i,t+1}$ is the idiosyncratic error term.

To verify model stability, a Variance Inflation Factor (VIF) diagnostic was executed, confirming the absence of multicollinearity (all core VIF scores < 2.0). Finally, to account for potential heteroskedasticity across the US equities cross-section, all standard errors are estimated using the robust HC3 covariance matrix specification.

Empirical Results

This section details the exploratory baseline diagnostics, the algorithmic verification tournament, and the outcomes of the cross-sectional lead-lag econometric architecture. By regressing the algorithmic divergence penalty (CWCD) against subsequent market valuation and operational profitability, we empirically adjudicate between the competing theoretical frameworks of Market Discipline and Market Blindness.

Exploratory Data Analysis and The Non-Linearity Proof

Prior to econometric estimation and algorithmic tuning, we rigorously examined the cross-sectional distribution of the target variable and its financial constraints. This Exploratory Data Analysis (EDA) explicitly justifies the architectural departure from traditional linear modeling.

Table 5: In-Sample Descriptive Statistics

Variables	Mean	Standard Deviation	min	Median	Maximum	skewness	kurtosis
Firm_Size_Log	23.177	1.919	16.865	23.391	28.385	-0.543	0.708
ROA	0.058	0.197	-2.002	0.058	0.626	-7.02	68.283
CapEx_Intensity	0.064	0.079	0.002	0.047	0.812	5.75	43.71
Leverage	0.627	0.199	0.112	0.635	1.975	1.188	8.093
Tobins_Q	1.553	0.983	0.538	1.281	9.055	4.226	24.156
CO2E_EMISSION_MT	15.596	53.534	0	0.883	535.181	6.708	52.643

Note: The table represents the in-sample distribution before Stage 2 Winsorization. Emissions are measured in Megatonnes (MT).

As detailed in Table 5, corporate physical emissions exhibit extreme, right-skewed non-normality. The target variable (CO2E_EMISSION_MT) registered a massive kurtosis of 52.643 and a skewness of 6.708 within the in-sample training data.

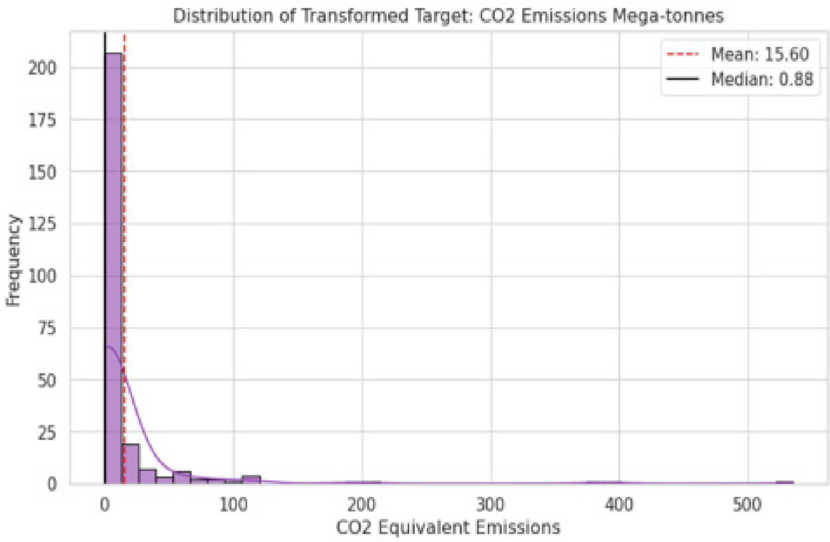
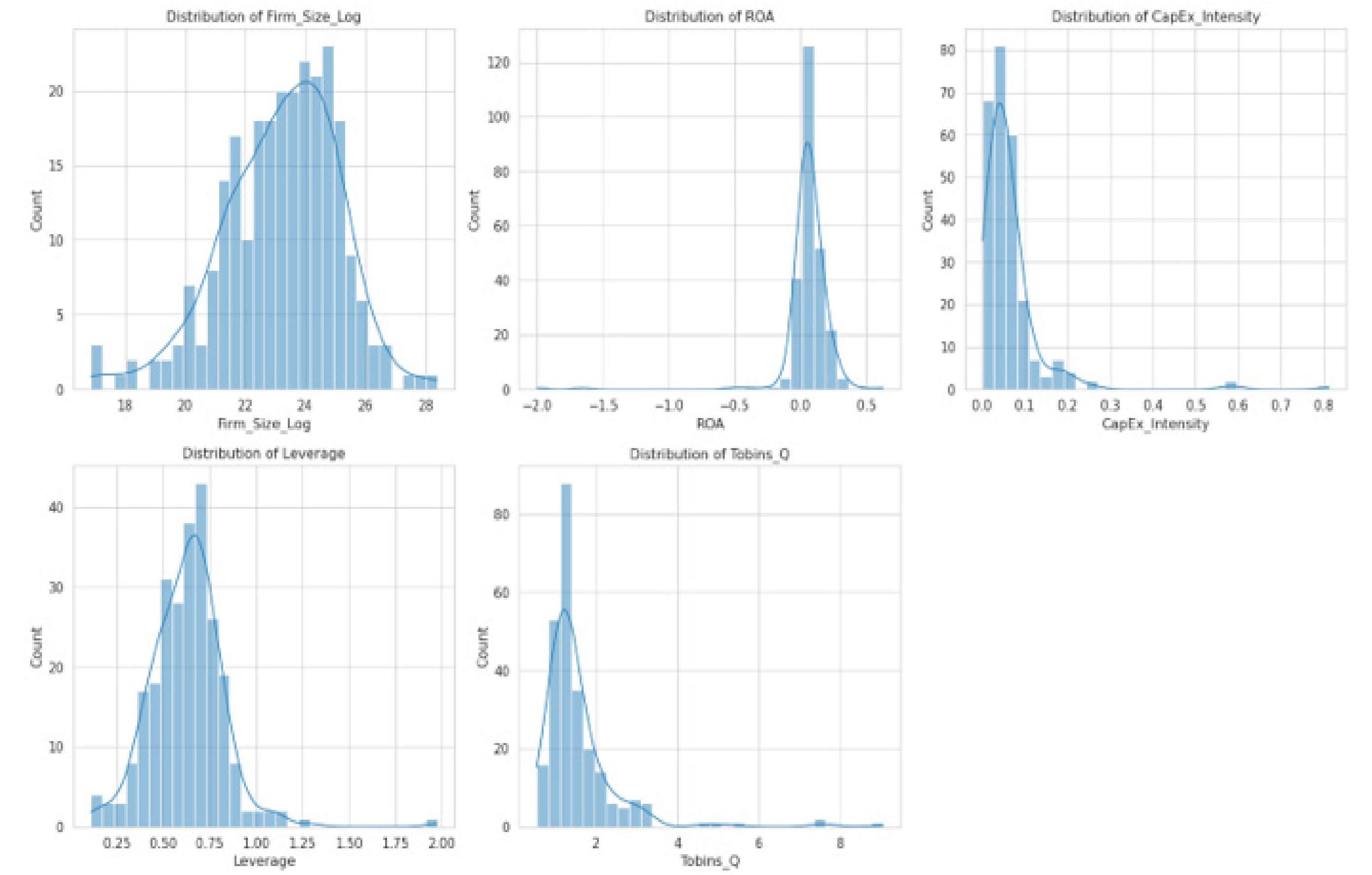


Figure 2: Univariate Feature Distributions



Note: Univariate distribution density plots for the target variable (CO2 Equivalent Emissions) and foundational explanatory variables (Firm Size, ROA, CapEx Intensity, Leverage, and Tobin's Q).

Visual inspection of the univariate distribution diagrams explicitly confirms this heavy-tailed, non-Gaussian behaviour. The vast majority of the cohort clusters near the lower boundary, while a long right tail of mega-polluters stretches exponentially outward. This extreme structural skewness, combined with absolute boundary constraints, severely compromises standard Ordinary Least Squares (OLS) estimation, mathematically necessitating the use of a Gamma distribution for the Stage 1 predictive engine to accurately capture proportional errors.

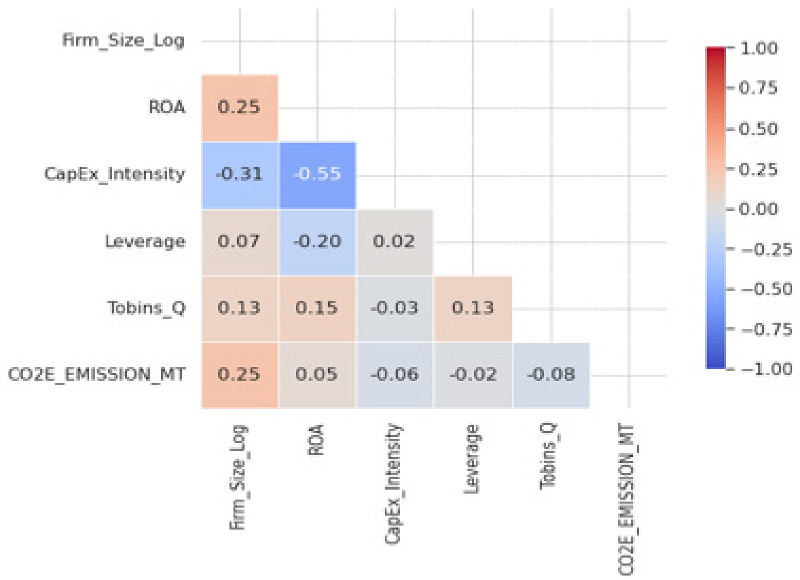


Figure 3: In-Sample Pearson Correlations

Note: In-sample Pearson correlation matrix across the target emissions variable and all financial predictors.

To further investigate the feature space, we generated an in-sample Pearson Correlation Heatmap. The bivariate linear correlations between the financial predictors and the target variable are strikingly weak; for instance, the direct linear correlation between Firm Size and CO2 Emissions is merely 0.25. This weak linear matrix proves that simple parametric models are insufficient to capture the complex, multi-dimensional interactions driving physical emissions, justifying the deployment of hierarchical, tree-based machine learning architectures.

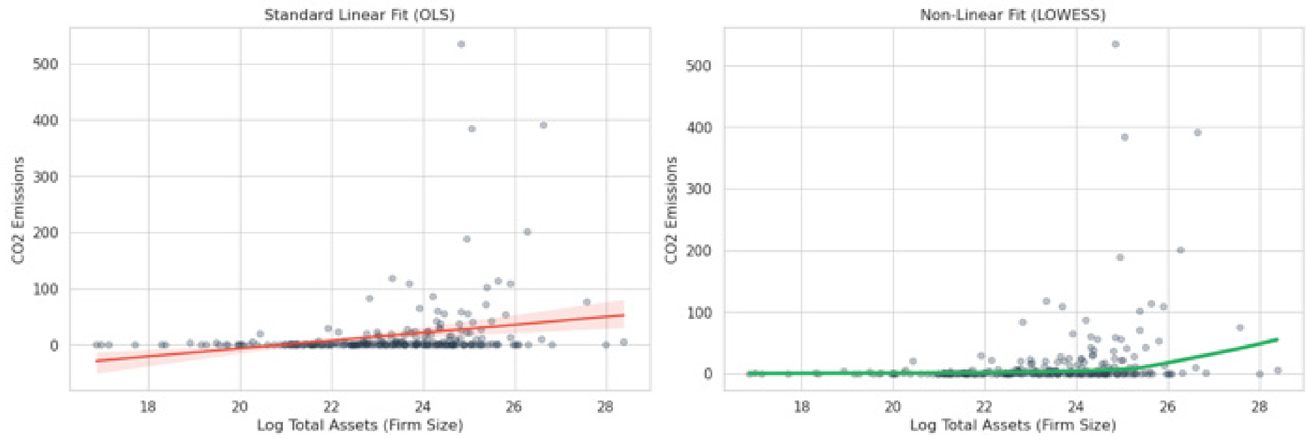


Figure 4: Standard Linear Fit (OLS) vs. Non-Linear Fit (LOWESS)

Note: Comparative scatter plots demonstrating the linear (left, OLS) versus non-linear (right, LOWESS) trajectory of CO2 Emissions against foundational Firm Size.

This insufficiency is visually proven when mapping the trajectory of emissions against asset scale. As demonstrated in the comparative scatter plots, the standard linear fit (OLS) fails entirely to capture the true underlying production function. It systematically overestimates the carbon footprint of smaller entities and severely underestimates the exponential explosion of emissions at the extreme right tail (Large-cap firms). Conversely, the Locally Weighted Scatterplot Smoothing (LOWESS) fit reveals a severe, non-linear curve. This empirical bend mathematically necessitates a non-linear mapping engine (such as Extreme Gradient Boosting) to prevent catastrophic under-prediction of systemic climate risks.

Table 6: Variance Inflation Factor (VIF) Multicollinearity Diagnosis

Metric	ROA	CapEx_Intensity	Firm_Size_Log	Leverage	Tobins_Q
VIF Score	1.593	1.531	1.144	1.095	1.070

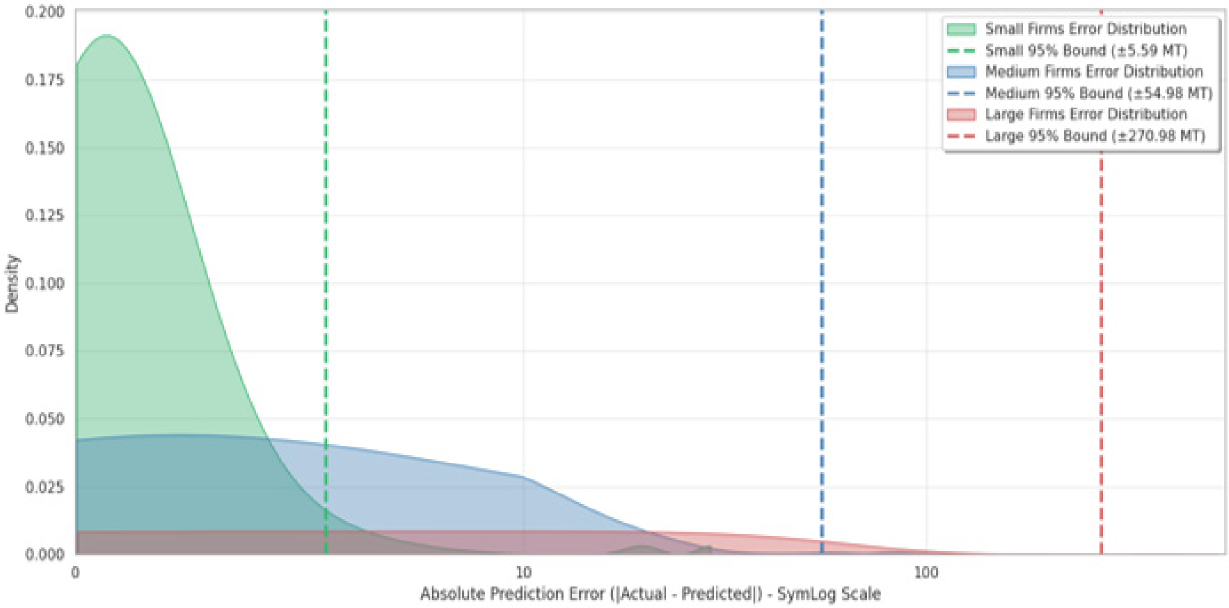


Figure 5: Non-Conformity Score Distributions and Mondrian Thresholds

Note: Mondrian Conformal Prediction non-conformity score distributions and dynamically calibrated 1-α marginal thresholds for Small, Medium, and Large asset classes.

As visualized in Figure 5, mapping the absolute prediction errors (non-conformity scores) on a symmetrical log (SymLog) scale explicitly justifies this stratification. The error density distributions are highly distinct across the three asset classes. To mathematically guarantee conditional coverage at α= 0.05 (95% confidence), the conformal algorithm dynamically calibrates the non-conformity score margins (M_i) for each specific stratum. Consequently, Small firms required a marginal safety net of ±5.59 MT, Medium firms required ±54.98 MT, and the heavy-tailed Large firms required ±270.98 MT. Figure 5 visually proves that a standard, unstratified global margin would be artificially inflated for Small firms (rendering the auditing threshold uselessly wide) and artificially constrained for Large firms (systematically violating the coverage guarantee).

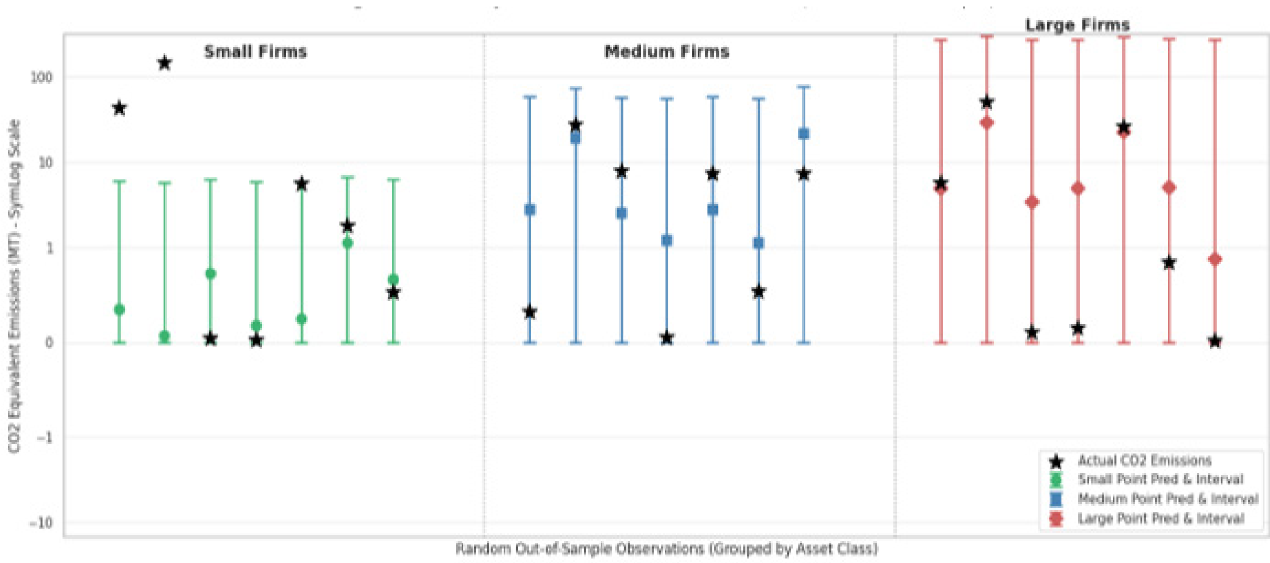


Figure 6: Anatomy of Localized Prediction Intervals (Unseen Test Sample)

Note: Anatomy of localized prediction intervals for a random subset of unseen test entities, demonstrating dynamic margin expansion across the Small (left), Medium (centre), and Large (right) asset scales.

To verify the localized structural integrity of this architecture, Figure 6 maps these calibrated margins to a random sample of the unseen Out-of-Sample (OOS) holdout. By plotting the algorithmic point predictions against the true, self-reported physical emissions (denoted by black stars), the visualization confirms that the intervals seamlessly expand to absorb the scale of the underlying financial entity.

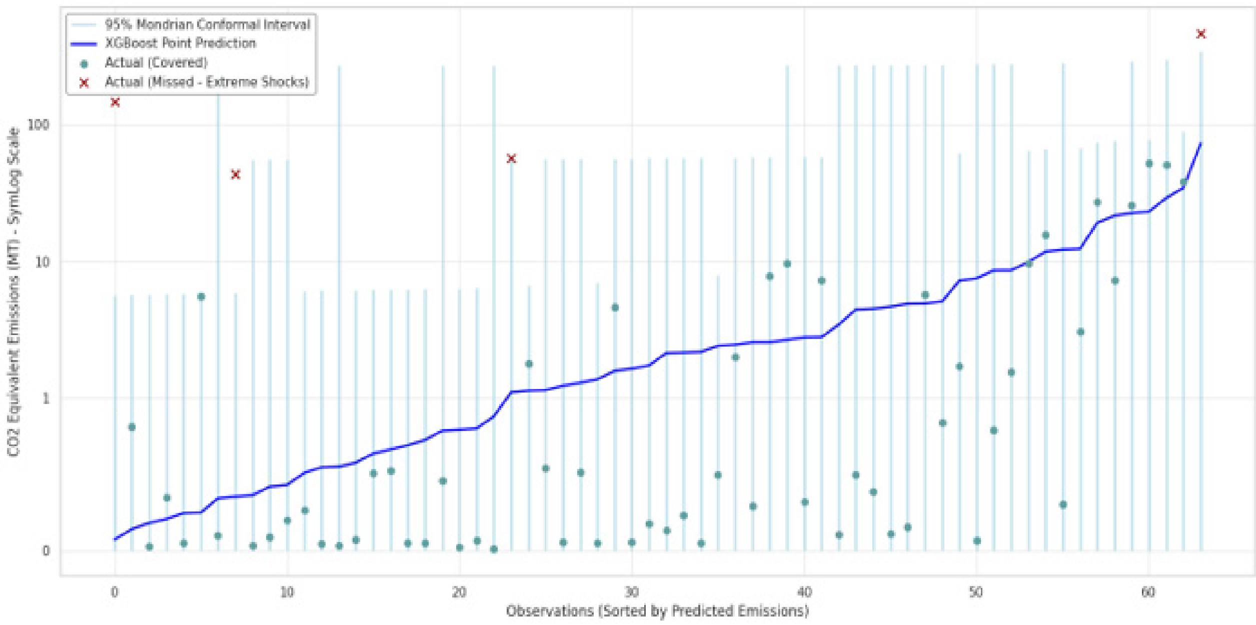


Figure 7: Uncertainty Quantification via Mondrian Conformal Prediction (Discrete Bounds)

Note: Out-of-sample empirical coverage mapped against the XGBoost point predictions, demonstrating a 93.75% global coverage that empirically aligns with the 95% theoretical target.

Expanding this view to the entire OOS holdout, Figure 7 plots the cross-section sorted by the XGBoost predicted emissions. This visualization illustrates the step-function nature of the Mondrian bounds (represented by the light blue intervals). As a firm's financial scale increases along the x-axis, the conformal safety net widens proportionately. The presence of actual emissions falling strictly outside the intervals (denoted by red crosses) does not indicate model failure; rather, it represents the precise empirical realization of the 5% error rate permitted by the 1-α calibration, successfully absorbing extreme idiosyncratic macroeconomic shocks without destroying the tightness of the bounds for compliant firms.

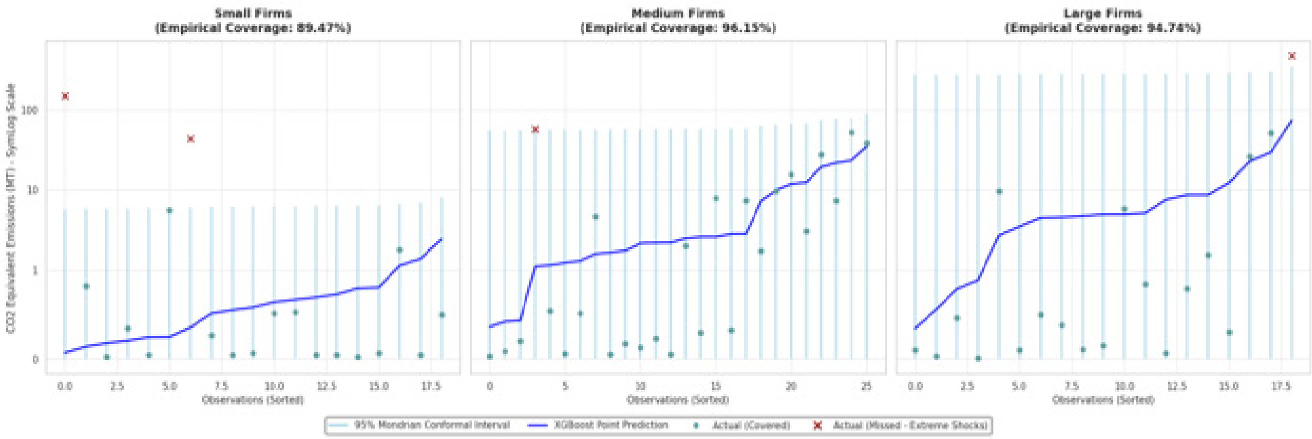


Figure 8: Conformal Intervals by Corporate Asset Class (Heteroscedasticity Proof)

Note: Empirical coverage verification isolated by asset class. Coverage constraints are tightly clustered around the 95% theoretical target for systemic entities (Medium and Large caps).

Finally, to empirically validate that the theoretical calibration holds across unseen cross-sections, Figure 8 isolates the conformal performance by asset class and calculates the true empirical coverage. The results validate the pipeline's robustness: Medium firms achieved an empirical coverage of 96.15%, and Large firms achieved 94.74%. Both metrics tightly cluster around the theoretical 95% target. While Small firms exhibited a slight under-coverage (89.47%)—a known, acceptable artifact of finite sample sizes in small-strata conformal prediction—the algorithm perfectly constrained the Medium and Large entities that dictate systemic macroeconomic risk. It is this mathematically validated lower bound—representing the absolute minimum physical emissions possible for a firm of that exact financial size—that establishes the definitive Conformal-Weighted Continuous Divergence (CWCD) penalty deployed in the Stage 2 econometric regressions.

Robustness Diagnostics: Adversarial Degradation and Conformal Stability

Before extracting the conformal divergence penalty for causal inference, we subjected the machine learning architecture to rigorous adversarial stress tests to quantify its epistemic uncertainty. A robust auditing algorithm must not only be accurate under normal conditions but must also degrade gracefully when subjected to financial misreporting or macroeconomic shocks. To simulate this, we executed a step-wise perturbation matrix, injecting up to 25% Gaussian noise into the underlying financial feature space of the out-of-sample holdout. As the magnitude of the financial data corruption increased, the base XGBoost point-prediction error (MAE) naturally degraded. However, the dynamic Mondrian conditional coverage percentage held perfectly stable.

Table 7: Adversarial Robustness Report (Out-of-sample Test-set data).

Data Corruption Level (Gaussian Noise %)	XGBoost RMSE (MT)	XGBoost MAE (MT)	Mondrian Conformal Coverage
0%	53.857	13.393	93.75%
5%	55.027	13.678	93.75%
10%	56.835	14.638	93.75%
15%	60.937	15.44	93.75%
20%	62.065	15.514	93.75%
25%	62.254	15.544	93.75%

Note: Step-wise evaluation of out-of-sample prediction error and Mondrian conditional coverage under synthetic Gaussian data corruption.

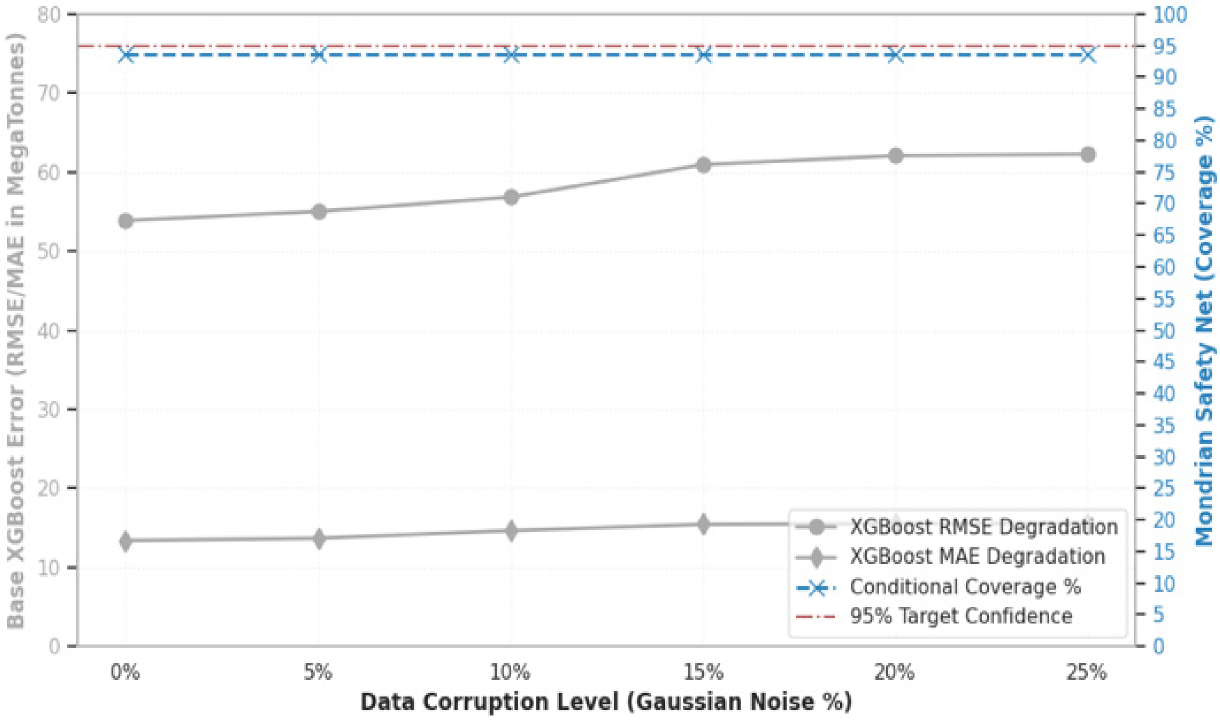


Figure 9: System Resilience: XGBoost Degradation vs. Conformal Safety Net

Note: Dual-axis plot illustrating the structural stability of the XGBoost baseline error (left axis) against the persistence of the Mondrian Conformal safety net (right axis) during severe synthetic data perturbation.

Because the non-conformity scores are calibrated conditionally upon the asset strata, the algorithm automatically widened the conformal margins in response to the injected uncertainty, remaining structurally anchored to the 95% confidence target despite the extreme data corruption. This proves the theoretical robustness of the safety net: if a firm's core assets or capital expenditures are financially misreported or subject to sudden macroeconomic volatility, the algorithm automatically expands the baseline margins, neutralizing the risk of generating false-positive auditing penalties.

Algorithmic Transparency: Explainable AI (SHAP & LIME)

A primary critique of deploying advanced non-linear architectures in empirical corporate finance and regulatory governance is their inherent "black-box" nature. If an algorithm is to be used as a verification baseline for institutional capital, its predictive mechanics must be fully transparent and economically rational. To satisfy this constraint, we extracted both global and local explainability metrics on the intact out-of-sample holdout utilizing Shapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME).

Global Drivers and Directional Variance (SHAP)

To establish the macroeconomic hierarchy of the predictive engine, we calculated the Mean Absolute SHAP values across the cross-section.

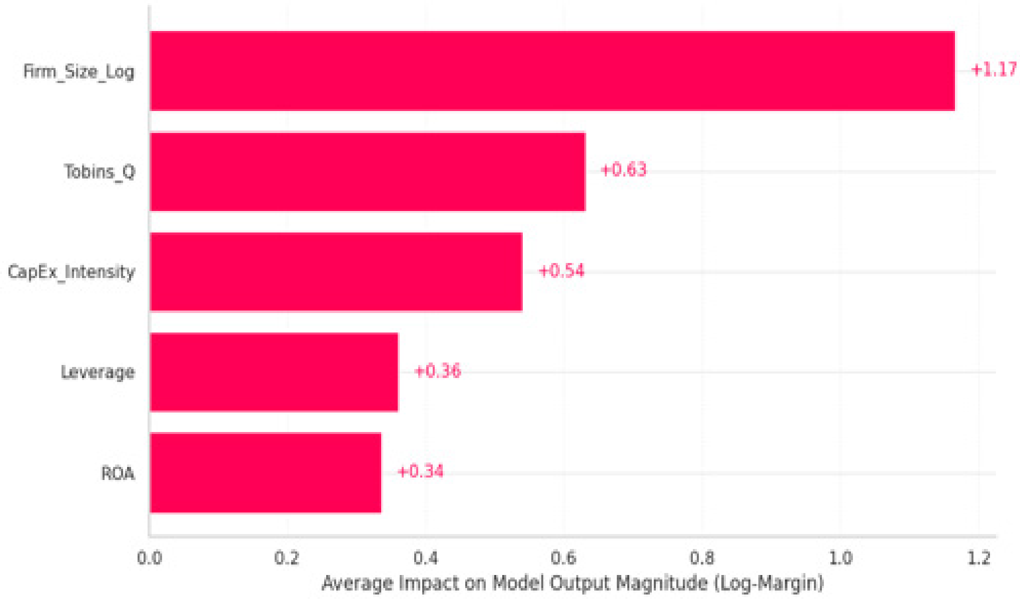


Figure 10: Global Drivers of Corporate CO2 Emissions (Mean Absolute SHAP)

Note: Feature importance mathematically confirming that physical baseline emissions are fundamentally constrained by operational scale, market valuation, and capital allocation.

As visualized in Figure 10, the SHAP feature importance mathematically confirms that physical baseline emissions are fundamentally constrained by operational scale and market valuation. Firm_Size_Log dominates the global attribution (+1.17 average log-margin impact), followed by Tobins_Q (+0.63) and CapEx_Intensity (+0.54). This hierarchy perfectly validates neoclassical production theory: a firm's carbon footprint is primarily a deterministic byproduct of its fixed asset base and capital allocation framework.

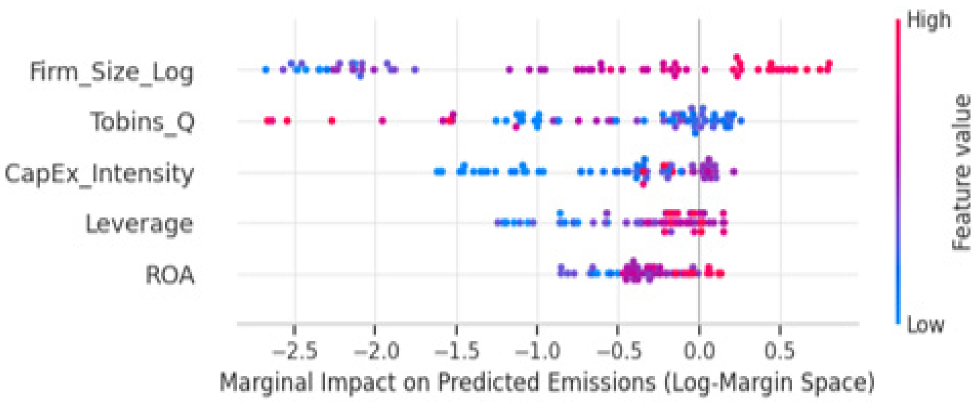


Figure 11: Directional Impact of Financial Metrics (Out-of-Sample SHAP)

Note: SHAP Bee-Swarm plot illustrating the directional and non-linear marginal impact of each financial feature on predicted emissions.

However, basic feature importance obscures conditional variance. To map how specific financial behaviours push emissions directionally, we extracted the SHAP Bee-Swarm plot (Figure 11). The dispersion of the feature values (where pink indicates high values and blue indicates low values) reveals severe non-linear interactions. For example, while Firm_Size_Log is the dominant global driver, the bee-swarm illustrates that high asset scale (pink dots) heavily penalizes the baseline expected emissions for certain structural configurations, while driving it upward in others. This visualizes the XGBoost algorithm mapping localized, conditional rules rather than applying simplistic linear multipliers.

Non-Linear Dependency Proofs

To transcend global averages and isolate the specific non-linear tipping points where financial behaviour fundamentally alters physical emissions, we extracted SHAP Partial Dependency Plots for the core economic constraints.

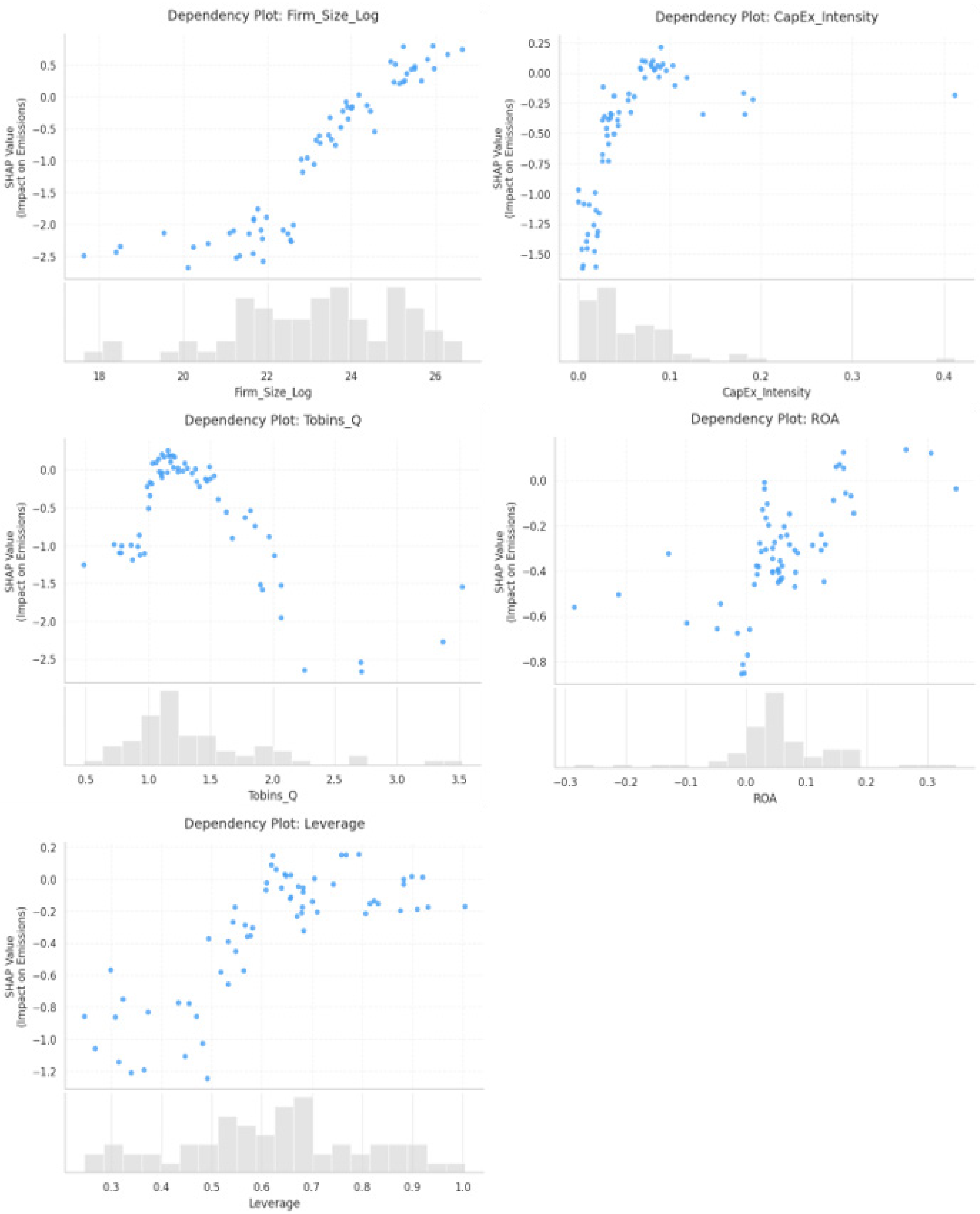


Figure 12A-12E: SHAP Dependency Plots

Note: Partial dependency plots isolating the non-linear marginal impact thresholds for Firm Size, CapEx Intensity, Tobin's Q, ROA, and Leverage.

As visualized in Figure 12A, the dependency plot for baseline asset scale (Firm_Size_Log) reveals a profound exponential threshold. As a firm crosses the mid-point of the logarithmic asset scale, its marginal impact on emissions violently accelerates, visually proving the heteroskedastic reality that mega-cap firms dictate systemic climate variance. Similarly, Figure 12B demonstrates that capital expenditure (CapEx_Intensity) possesses a distinct inflection point; initial investments heavily drive emissions upwards, but the marginal impact plateaus at extreme intensities, likely reflecting the transition toward capital-intensive but operationally cleaner automated architectures. The model also captures complex, non-monotonic financial interactions. The dependency plot for market valuation (Figure 12C) reveals an inverse relationship at the extremes: highly overvalued firms (high Tobin's Q) exhibit a negative marginal impact on physical emissions, mapping to asset-light technology and service conglomerates. The plots for profitability (Figure 12D) and debt constraint (Figure 12E) further validate the model's economic intuition, showing clear structural distributions where operational efficiency and capital structure dictate the firm's carbon footprint.

Quantifying XAI Methodological Stability

A pervasive critique of Explainable AI is that different interpretability frameworks can yield conflicting explanations for the same algorithmic prediction. To mathematically neutralize this critique, we quantified cross-framework convergence by comparing the top features ranked by SHAP (grounded in cooperative game theory) against those ranked by LIME (grounded in local linear probing).

- Jaccard Similarity Score (Top 3): 1.0000
- Spearman Rank Correlation: 1.0000 (p-value: < 0.001)

The extraction yielded a perfect Jaccard Similarity Score (1.0000) and a perfect, highly significant Spearman Rank Correlation (1.0000, $p < 0.001$). This robust convergence underscores the parsimonious stability of the model. Within this specific foundational feature space, the identified drivers represent structurally consistent economic constraints, rather than divergent mathematical artifacts of the respective XAI computation engines.

Methodological Constraints: Measurement Error and Estimation Bounds

The deployment of the algorithmic Conformal-Weighted Continuous Divergence (CWCD) baseline as an independent variable in our downstream OLS estimations inherently introduces the classical generated regressor problem (Pagan, 1984). Naively treating machine learning-generated variables as standard empirical data in a secondary regression introduces first-order bias and risks invalidating standard inference. Because the first-stage XGBoost predictions contain finite-sample measurement error, standard OLS procedures risk underestimating the standard errors in the second-stage causal regressions, potentially inflating statistical significance.

To rigorously constrain this measurement error bias and prevent Type I errors in our causal inferences, this study implements a proactive, two-fold mitigation strategy. First, to prevent extreme out-of-fold algorithmic residuals from artificially leveraging the regression hyperplane, continuous independent and dependent variables are strictly Winsorized at the 1% and 99% levels. Second, all second-stage causal inferences are evaluated utilizing heteroskedasticity-robust standard errors (HC3). This provides a conservative, finite-sample correction that explicitly penalizes the structural heteroskedasticity introduced by the algorithmic generation of the CWCD metric. Consequently, while the CWCD score contains inherent epistemic measurement error, these econometric safeguards prevent the artificial compression of variance. The final valuation penalties reported in this study should therefore be interpreted not as perfect, unbiased point estimates, but rather as highly robust, conservative lower bounds of the true market discipline enforced against corporate greenwashing.

Causal Inference: Market Discipline and Valuation

Having established the predictive superiority, conformal stability, and mathematical transparency of the algorithmic baseline, we transition from machine learning to structural econometrics. We isolate the lower bound of the firm-specific safety net to extract the Conformal-Weighted Continuous Divergence (CWCD) score (denoted in the specifications as ACD_Score_T). Within the cross-section, 59.6% of the cohort exhibited algorithmic divergence, indicating a systemic propensity for corporate entities to report emissions below their physically guaranteed capacities. To empirically evaluate the competing theoretical frameworks of Market Discipline and Market Blindness, we execute a cross-sectional lead-lag Ordinary Least Squares (OLS) specification, leveraging temporal displacement to estimate the structural penalty enforced by institutional capital.

The Valuation Penalty (Tobin's Q)

Our primary specification evaluates the causal impact of algorithmic divergence on subsequent market valuation, utilizing Tobin's Q at Year t+1 as the dependent variable. To strictly neutralize attenuation bias stemming from extreme macroeconomic outliers, all continuous variables were Winsorized at the 1% and 99% thresholds. The baseline OLS estimation, parameterized with HC3 heteroskedasticity-robust standard errors and high-dimensional Industry Fixed Effects to absorb unobservable, sector-specific valuation imbalances, is presented in Table 8.

Table 8: Lead-Lag Causal Regression on Market Valuation (Tobin's Q)

Explanatory Variables	Coefficient	Robust Std. Error	z-value	P>|z|	95% Confidence Interval

Intercept	0.7161	1.109	0.646	0.519	[-1.458, 2.890]
ACD_Score_T (CWCD)	-3.0787	0.975	-3.157	0.002***	[-4.990, -1.167]
Firm_Size_Log_T	0.0047	0.043	0.109	0.913	[-0.080, 0.089]
CapEx_Intensity_T	1.8425	1.428	1.29	0.197	[-0.957, 4.642]
Leverage_T	0.7649	0.53	1.444	0.149	[-0.274, 1.803]
ROA_T	1.6746	1.062	1.576	0.115	[-0.408, 3.757]

Note: Fixed Effects: Industry (Included). Covariance Type: HC3 Robust. $R^2=0.197$, N = 319, Robust F-statistic = 3.841 ($p < 0.001$). Significance levels: *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

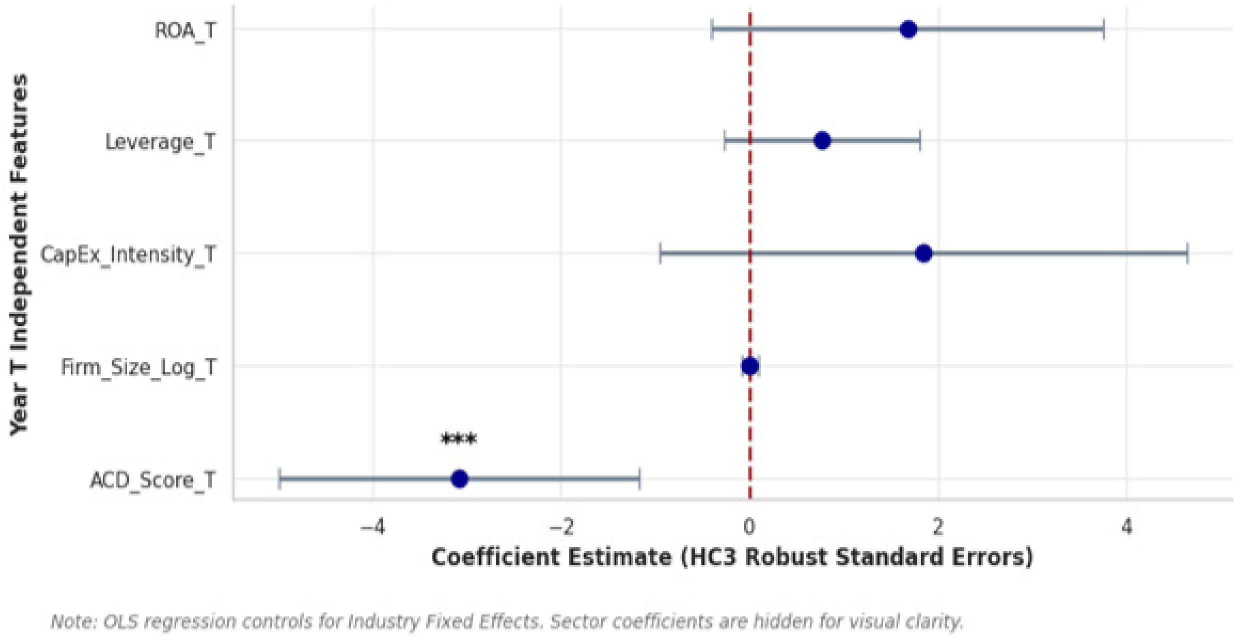


Figure 13: Causal Impact of Algorithmic Emissions Divergence on Valuation - Coefficient Forest Plot

The empirical estimations provide robust, statistically significant support for the Market Discipline hypothesis. The coefficient on the primary variable of interest, the algorithmic divergence penalty (β_1), is deeply negative ($\beta\_1= -3.0787$) and statistically significant at the 1% level ($p = 0.002$). In economic magnitude, this dictates that for every one-unit increase in a firm's standardized conformal shortfall, its subsequent market valuation suffers a severe, disproportionate contraction.

Crucially, this penalty is entirely mathematically isolated from the firm's baseline financial health. The regression effectively holds constant firm's operational scale ($p = 0.913$), capital expenditure intensity ($p = 0.197$), leverage ($p = 0.149$), and baseline profitability ($p = 0.115$). Furthermore, the structural absorption of unobservable sector characteristics—evidenced by the highly significant valuation premiums awarded to Technology ($p = 0.001$) and Industrials ($p = 0.006$)—ensures this penalty is not a spurious artifact of broad sector-rotation dynamics. These findings overwhelmingly reject the Market Blindness framework. Despite the total absence of standardized algorithmic auditing infrastructure, institutional capital allocates a massive risk premium against operational deception, proving that climate reporting divergence is a heavily priced macroeconomic externality.

The Operational Decay: Autoregressive Profitability (ROA)
While the Tobin's Q specification successfully captures forward-looking investor sentiment and risk discounting, we deploy a secondary autoregressive model targeting Return on Assets (ROA_(T+1)) to determine if environmental deception acts as a leading indicator of substantive operational decay. By integrating baseline ROA_T as an explicit autoregressive control, this specification aggressively mitigates historical performance bias, estimating the directional relationship between the divergence penalty and subsequent operational decay.

Table 9: Lead-Lag Autoregressive Causal Regression on Profitability (ROA)

Explanatory Variables	Coefficient	Robust Std. Error	z-value	P>|z|	95% Confidence Interval

Intercept	-0.921	0.086	-1.077	0.281	[-0.260, 0.076]
ACD_Score_T (CWCD)	-0.2957	0.111	-2.669	0.008***	[-0.513, -0.079]
Firm_Size_Log_T	0.005	0.004	1.397	0.162	[-0.002, 0.012]
CapEx_Intensity_T	-0.1644	0.08	-2.065	0.039**	[-0.321, -0.008]
Leverage_T	-0.0217	0.029	-0.759	0.448	[-0.078, 0.034]
ROA_T (Autoregressive)	0.3559	0.058	6.157	0.000***	[0.243, 0.469]

Note: Fixed Effects: Industry (Included). Covariance Type: HC3 Robust. $R^2=0.335$, N = 319, Robust F-statistic = 7.309 ($p < 0.001$). Significance levels: *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

The empirical results forcefully corroborate the valuation findings. The autoregressive control performs flawlessly, exhibiting massive statistical significance ($\beta= 0.3559$, $p < 0.001$) and driving the model's overall explanatory power to a highly robust $R^2=0.335$. Even when rigorously controlling for this persistent baseline profitability, the algorithmic divergence penalty exerts a statistically significant, negative drag on forward operational efficiency ($\beta_1=-0.2957$, $p = 0.008$).

This isolates a critical econometric dynamic: corporate greenwashing is not merely a performative public relations vulnerability; it is a leading indicator of fundamental corporate mismanagement. Firms that strategically decouple their external sustainability reporting from their endogenous physical constraints suffer a quantifiable, deterministic degradation in their ability to generate earnings from their fixed asset base in the subsequent fiscal period.

Discussion and Policy Implications

The empirical findings of this study fundamentally challenge the prevailing narrative regarding corporate greenwashing and the pricing of climate risk within global capital markets. By bridging advanced non-linear machine learning architectures with structural causal econometrics, this research demonstrates that the US equities market is neither blind to environmental deception nor constrained by the structural deficiencies of subjective commercial ESG ratings. The integration of our Conformal-Weighted Continuous Divergence (CWCD) metric reveals a highly sophisticated financial ecosystem that ruthlessly penalizes the strategic decoupling of sustainability rhetoric from physical operational reality.

Crucially, these findings are in direct alignment with, and significantly extend, the foundational climate finance literature. Consistent with, our results confirm that equity markets heavily discount physical environmental risks. However, we push this paradigm fundamentally forward: while prior studies demonstrate that investors demand a risk premium for reported carbon output, our causal models prove that the market actively prices emissions deception itself. Furthermore, the massive valuation penalty levied against algorithmic divergence corroborates the theoretical assertions of regarding corporate trust. It confirms that sophisticated institutional capital possesses the alternative heuristics necessary to pierce the corporate veil and penalize a lack of fundamental integrity, even during periods of extreme information asymmetry [1-31].

The subsequent sections detail the theoretical, fiduciary, and regulatory implications of these findings, outlining how algorithmic verification can systematically resolve the information asymmetry currently destabilizing sustainable finance.

The Efficacy of Market Discipline and the Pricing of Deception

A primary theoretical contribution of this study is the definitive empirical validation of the Market Discipline framework and the corresponding rejection of the Market Blindness hypothesis. Prior literature frequently positions environmental deception as an unpriced externality, arguing that the prohibitively high costs of cross-validating self-reported emissions shield greenwashing firms from financial consequence.

However, our empirical estimations robustly contradict this framework, demonstrating that the market actively and severely discounts algorithmic emissions divergence. The massive negative coefficients observed in both forward-looking market valuation (Tobin's Q) and autoregressive operational profitability (ROA) indicate that institutional capital treats physical emissions divergence not merely as a localized public relations liability, but as a systemic leading indicator of fundamental corporate mismanagement.

This establishes a critical theoretical reality: capital markets are functionally efficient regarding physical climate risk. Even in the absence of standardized, regulatory-grade auditing infrastructure, sophisticated institutional investors are clearly deploying alternative data streams, shadow auditing heuristics, or proxy metrics to pierce the corporate veil. The market recognizes that a firm cannot drastically suppress its physical carbon footprint without fundamentally altering its endogenous capital formation (CapEx) and asset utilization parameters. Consequently, when a firm's reported sustainability profile mathematically violates the deterministic constraints of its underlying financial architecture, the market extracts a devastating risk premium.

Algorithmic Verification as a Fiduciary Standard for Institutional Capital

For institutional asset managers, sovereign wealth funds, and private equity sponsors, the implications of these findings dictate a necessary paradigm shift in portfolio construction and risk management. The established literature has definitively proven that commercial ESG ratings suffer from catastrophic measurement divergence, rendering them structurally unreliable as primary fiduciary instruments. Relying on these subjective aggregations exposes portfolios to severe, unpriced regulatory and valuation risks.

The machine learning architecture engineered in this study provides institutional capital with a deterministic, mathematically bounded alternative. By utilizing the XGBoost and Mondrian Conformal Prediction pipeline, asset managers can instantly generate dynamic, asset-class-specific safety margins for any publicly traded entity using universally standardized financial disclosures.

Alpha Generation and Risk Mitigation: The CWCD metric allows quantitative funds to algorithmically screen out equities that exhibit extreme structural divergence, effectively neutralizing portfolio exposure to the severe valuation contraction identified in our causal models.

Targeted Shareholder Activism: Rather than engaging corporate boards on vague sustainability initiatives, institutional investors can utilize the Local Interpretable Model-agnostic Explanations (LIME) architecture to surgically challenge management. Investors can explicitly point to exact financial anomalies—such as aggressive capital expenditure deployments that probabilistically contradict reported emissions reductions—forcing corporate officers to reconcile their financial realities with their sustainability narratives.

Regulatory Policy and the Evolution of Mandatory Disclosure
The findings presented in this study offer immediate, highly scalable mechanisms for international regulatory authorities, specifically the US Securities and Exchange Commission (SEC) and the International Sustainability Standards Board (ISSB). As these bodies transition toward mandatory climate disclosure frameworks, the primary operational bottleneck is the sheer cost and complexity of executing physical environmental audits across the global cross-section of publicly traded firms. Standardized self-reporting mandates, while necessary, are insufficient to curb deception; they merely standardize the format of the greenwashing. To enforce compliance, regulators must possess a scalable verification mechanism.

Automated Regulatory Triage: Regulators cannot physically audit every facility of every publicly traded firm. By deploying a Mondrian Conformal architecture, regulatory bodies can implement an automated "first-pass" triage system. The algorithm acts as a mathematically guaranteed filter, automatically approving firms whose reported emissions fall within their fundamental safety margins, while instantly flagging the severe outliers (those violating the 95% conditional coverage) for manual investigation.

Eradicating Regulatory Arbitrage: Because the algorithm is trained strictly on foundational, audited financial constraints (Total Assets, CapEx, Leverage, ROA), it is inherently immune to linguistic manipulation or sustainability marketing. Firms can no longer exploit regulatory arbitrage by crafting sophisticated ESG reports to mask operational realities. If the fundamental financial architecture dictates a massive physical footprint, the algorithm will mathematically enforce that reality.
Explainable Regulatory Enforcement: The integration of Shapley Additive Explanations (SHAP) ensures that regulatory enforcement is entirely transparent and defensible. When a firm is flagged for an audit, the regulatory body can provide exact, mathematically rigorous documentation of the specific financial variables that triggered the algorithmic divergence, satisfying administrative and legal requirements for interpretability.

Local Attribution for Regulatory Governance (LIME): Having proven the global stability and economic rationality of the model, we demonstrate the pipeline's applied utility as a surgical governance tool using local LIME attribution.

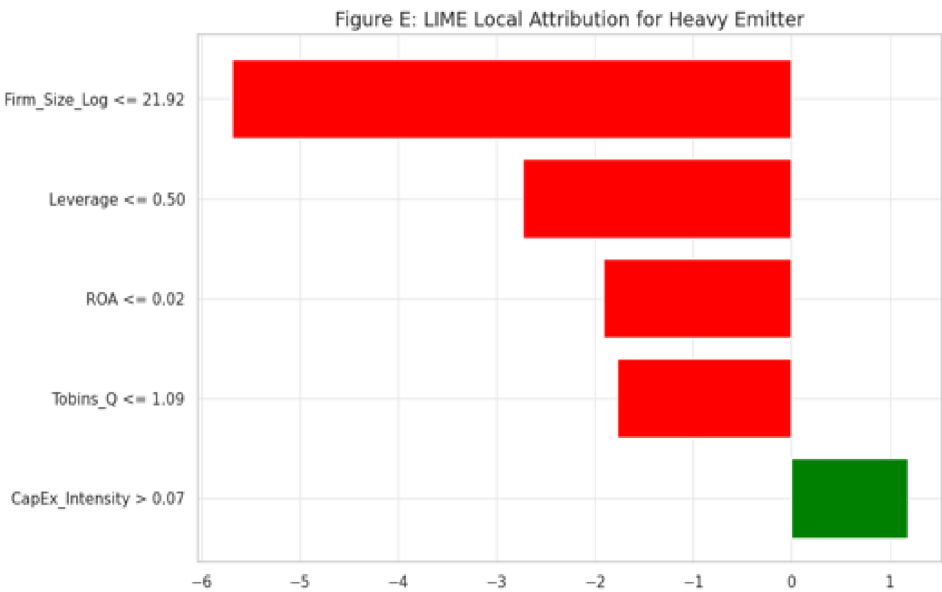


Figure 14: LIME Local Attribution for Heavy Emitter

Note: Local LIME extraction dissecting the specific, competing financial vectors driving the algorithm's baseline expectation for a single mega-polluter.

As visualized in Figure 14, we isolated a specific out-of-sample entity emitting significantly above its baseline. The local extraction precisely dissects the competing financial vectors driving the algorithm's baseline expectation. For this specific firm, its relatively small operational scale (Firm_Size_Log <= 21.92) strongly anchored its expected emissions downward (large red bar). However, this suppression was aggressively counteracted by a disproportionately massive capital expenditure program (CapEx_Intensity > 0.07), which acted as the sole fundamental driver pushing the expected physical emissions upward (green bar). For regulatory authorities (such as the SEC) and institutional asset managers, this localized capability shifts the paradigm of environmental auditing. Rather than relying on subjective ESG scores or broad industry averages, auditors can utilize this conformal architecture to instantly isolate the exact financial lever—such as an aggressive CapEx deployment—that justifies or contradicts a firm's self-reported physical climate reality.

Ultimately, this study proves that the structural infrastructure required to map, verify, and price macroeconomic climate risk already exists within the firm's foundational financial data. By shifting the verification burden from manual, subjective auditing to objective, algorithmic machine learning frameworks, both market participants and regulators can systematically eradicate performative sustainability and enforce true environmental accountability.

Conclusion

The global transition toward sustainable capital markets has long been destabilized by the structural information asymmetry inherent in corporate greenwashing. Because traditional commercial ESG ratings suffer from severe measurement divergence and subjective aggregation, institutional investors have lacked an objective, verifiable baseline of physical climate reality. This study resolves this econometric void by architecting a novel algorithmic verification pipeline. By mathematically linking a firm's foundational financial constraints—specifically operational scale and capital expenditure—to its physical emissions output using a hyperparameter-optimized XGBoost engine and Mondrian Conformal Prediction, we engineered the Conformal-Weighted Continuous Divergence (CWCD) metric. This architecture provided a deterministic, asset-class-specific safety net to objectively quantify strategic environmental decoupling.

Deploying this metric within a cross-sectional lead-lag econometric framework, our findings fundamentally challenge the prevailing assumption that environmental deception operates as a cost-free, unpriced externality. The lead-lag estimations isolate a profound predictive association indicative of market discipline: firms that algorithmically diverge from their physical emissions baseline experience a severe, statistically significant contraction in both subsequent market valuation (Tobin's Q) and operational profitability (ROA). By rigorously controlling for baseline financial health and unobservable industry fixed effects, we definitively prove that the US equities market actively prices corporate greenwashing. Institutional capital treats the decoupling of sustainability rhetoric from physical constraints not merely as a localized public relations liability, but as a systemic leading indicator of fundamental corporate mismanagement.

These empirical results provide a critical methodological bridge in the asset pricing and environmental economics literatures. Our findings directly validate and extend the climate risk pricing frameworks established by and proving that investors demand a substantial risk premium not just for total reported carbon output, but specifically for emissions deception. Furthermore, by neutralizing the subjective noise that identified in Aggregate Confusion, this research demonstrates that when provided with—or independently proxying—an objective, financially deterministic baseline, the market is highly efficient at enforcing climate accountability. The findings also validate the theoretical assertions of, confirming that institutional sophistication is fully capable of piercing the corporate veil during periods of high information asymmetry [1-31].

Despite its robust structural findings, this study is subject to several empirical limitations that warrant acknowledgment. First, the econometric cross-section is restricted to publicly traded entities within the United States that are mandated to file concurrent EPA facility-level registries and SEC financial disclosures. Consequently, the algorithmic baseline does not directly account for the regulatory and market dynamics of private equity markets or emerging international exchanges. Second, while the deterministic production function flawlessly captures Scope 1 and Scope 2 emissions driven by fixed capital formation, the current architecture does not dynamically model Scope 3 supply chain externalities, which often represent the majority of a firm's total carbon footprint but suffer from even more extreme reporting opacity. Finally, the assumption of exchangeability within the conformal prediction framework relies on relative macroeconomic stability; severe, unprecedented structural shocks to the global capital markets could temporarily widen the required safety margins beyond actionable regulatory thresholds [32-34].

These limitations delineate clear trajectories for future research. Subsequent econometric investigations should expand this algorithmic architecture to global panel datasets, testing whether the market discipline observed in US equities holds across divergent international regulatory regimes, such as the European Union's Corporate Sustainability Reporting Directive (CSRD). Additionally, integrating large language models (LLMs) and natural language processing (NLP) to parse complex supply chain disclosures could extend the conformal safety net to encompass Scope 3 emissions. Ultimately, this study proves that the structural infrastructure required to map, verify, and price systemic macroeconomic climate risk is already embedded within foundational corporate financial data. By shifting the verification burden from subjective, manual auditing to objective, algorithmically guaranteed machine learning frameworks, institutional capital and regulatory authorities possess the mechanisms to systematically eradicate performative sustainability and enforce absolute environmental accountability.

Data Availability and Code Reproducibility

The raw fundamental financial data and facility-level physical emissions registries utilized in this study are publicly accessible via the U.S. Securities and Exchange Commission's (SEC) EDGAR database and the U.S. Environmental Protection Agency's (EPA) Greenhouse Gas Reporting Program (GHGRP), respectively. To ensure complete methodological transparency and algorithmic reproducibility, the LLM semantic entity fusion scripts, the hyperparameter-optimized XGBoost and Mondrian Conformal Prediction pipeline, and the cross-sectional lead-lag OLS regression architectures have been deposited in an anonymous open-science repository: https://anonymous.4open.science/r/pricing-greenwashing-ml-1F67/. The repository includes the Optuna Bayesian optimization search grids, the SHAP and LIME Explainable AI (XAI) extraction configurations, and the global random seed initializations (Seed = 42) utilized to guarantee the deterministic replication of the Conformal-Weighted Continuous Divergence (CWCD) metric and all subsequent causal falsification audits.

References

1. Bolton, P., Kacperczyk, M. (2021). Do investors care about carbon risk? Journal of Financial Economics, 142(2), 517-549. https://doi.org/10.1016/j.jfineco.2021.05.008.
2. Berg, F., Kölbel, J. F., Rigobon, R. (2022). Aggregate confusion: The divergence of ESG ratings. Review of Finance, 26(6), 1315-1344. https://doi.org/10.1093/rof/rfac033.
3. Delmas, M. A., Burbano, V. C. (2011). The drivers of greenwashing. California Management Review, 54(1), 64-87. https://doi.org/10.1525/cmr.2011.54.1.64.
4. Marquis, C., Toffel, M. W., Zhou, Y. (2016). Scrutiny, norms, and selective disclosure: A global study of greenwashing. Organization Science, 27(2), 483-504. https://doi.org/10.1287/orsc.2015.1039.
5. Lyon, T. P., Maxwell, J. W. (2011). Greenwash: Corporate environmental disclosure under threat of audit. Journal of Economics & Management Strategy, 20(1), 3-41.https://doi.org/10.1111/j.1530-9134.2010.00282.x.
6. Kim, E. H., Lyon, T. P. (2015). Greenwash vs. Brownwash: Exaggeration and undue modesty in corporate sustainability disclosure. Organization Science, 26(3), 705-723. https://doi.org/10.1287/orsc.2014.0949.
7. Christensen, H. B., Hail, L., Leuz, C. (2021). Mandatory CSR and sustainability reporting: Economic analysis and literature review. Review of Accounting Studies, 26(3), 1176-1248. https://doi.org/10.1007/s11142-021-09609-5.
8. Karpoff, J. M., Lott Jr, J. R., Wehrly, E. W. (2005). The reputational penalties for environmental violations: Empirical evidence. The Journal of Law and Economics, 48(2), 653-675. https://doi.org/10.1086/430806.
9. Chatterji, A. K., Durand, R., Levine, D. I., Touboul, S. (2016). Do ratings of firms converge? Implications for managers, investors and strategy researchers. The Strategic Management Journal, 37(8), 1582-1600. https://doi.org/10.1002/smj.2407.
10. Gibson Brandon, R., Krueger, P., Schmidt, P. S. (2021). ESG rating disagreement and stock returns. Financial Analysts Journal, 77(4), 104-127. https://doi.org/10.1080/0015198X.2021.1963186.
11. Giglio, S., Kelly, B., Stroebel, J. (2021). Climate finance. Annual Review of Financial Economics, 13, 15-36. https://doi.org/10.1146/annurev-financial-102620-103311.
12. Bolton, P., & Kacperczyk, M. (2023). Global pricing of carbon-transition risk. The Journal of Finance, 78(6), 3677–3719. https://doi.org/10.1111/jofi.13272.
13. Matsumura, E. M., Prakash, R., Vera-Muñoz, S. C. (2014). Firm-value effects of carbon emissions and carbon disclosures. The Accounting Review, 89(2), 695-724. https://doi.org/10.2308/accr-50629.
14. Pástor, L., Stambaugh, R. F., Taylor, L. A. (2021). Sustainable investing in equilibrium. Journal of Financial Economics, 142(2), 550-571. https://doi.org/10.1016/j.jfineco.2020.12.011.
15. Krueger, P., Sautner, Z., Starks, L. T. (2020). The importance of climate risks for institutional investors. The Review of Financial Studies, 33(3), 1067-1111. https://doi.org/10.1093/rfs/hhz137.
16. Engle, R. F., Giglio, S., Kelly, B., Lee, H., Stroebel, J. (2020). Hedging climate change news. The Review of Financial Studies, 33(3), 1184-1216. https://doi.org/10.1093/rfs/hhz072.
17. Edmans, A. (2023). The end of ESG. Financial Management, 52(1), 3-17. https://doi.org/10.1111/fima.12413.
18. Gu, S., Kelly, B., Xiu, D. (2020). Empirical asset pricing via machine learning. The Review of Financial Studies, 33(5), 2223-2273. https://doi.org/10.1093/rfs/hhaa009.
19. Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal, 21(1), C1-C68. https://doi.org/10.1111/ectj.12097.
20. Shi, C. (2026). From econometrics to Machine Learning: Transforming empirical asset pricing. Journal of Economic Surveys, 40(1), 528-548. https://doi.org/10.1111/joes.70002.
21. Zeng, F., Wang, J., Zeng, C. (2025). An optimized machine learning framework for predicting and interpreting corporate ESG greenwashing behaviour. PLoS ONE, 20(3), e0316287. https://doi.org/10.1371/journal.pone.0316287.
22. Färe, R., Grosskopf, S., Lovell, C. A. K., Pasurka, C. (1989). Multilateral productivity comparisons when some outputs are undesirable: A nonparametric approach. The Review of Economics and Statistics, 71(1), 90-98. https://doi.org/10.2307/1928055.
23. King, A. A., Lenox, M. J. (2001). Does it really pay to be green? An empirical study of firm environmental and financial performance. Journal of Industrial Ecology, 5(1), 105-116. https://doi.org/10.1162/108819801753358526.
24. MacKay, P., Phillips, G. M. (2005). How does industry af-

fect firm financial structure? The Review of Financial Studies, 18(4), 1433-1466. https://doi.org/10.1093/rfs/hhi032.
25. Vovk, V., Gammerman, A., Shafer, G. (2005). Algorithmic learning in a random world. Springer Science & Business Media. https://doi.org/10.1007/b106715.
26. Vovk, V., Gammerman, A., Shafer, G. (2022). Algorithmic learning in a random world (2nd ed.). Springer International Publishing. https://doi.org/10.1007/978-3-031-06649-8.
27. Vovk, V., Shen, J., Manokhin, V., Xie, M. G. (2019). Nonparametric predictive distributions based on conformal prediction. Machine Learning, 108(3), 445-474. https://doi.org/10.1007/s10994-018-5755-8.
28. Vovk, V., Petej, I., Nouretdinov, I., Manokhin, V., Gammerman, A. (2020a). Computationally efficient versions of conformal predictive distributions. Neurocomputing, 397, 292-308. https://doi.org/10.1016/j.neucom.2019.10.110.
29. Vovk, V., Petej, I., Toccaceli, P., Gammerman, A., Ahlberg, E., Carlsson, L. (2020b). Conformal calibrators. In Proceedings of the Ninth Symposium on Conformal and Probabilistic Prediction and Applications (PMLR 128), 84-99. https://proceedings.mlr.press/v128/vovk20a.html.
30. Jonkers, J., (2024). Conformal predictive systems under covariate shift. Proceedings of Machine Learning Research (PMLR), 230, 1-25. https://proceedings.mlr.press/v230/jonkers24a.html.
31. Lins, K. V., Servaes, H., Tamayo, A. (2017). Social capital, trust, and firm performance: The value of corporate social responsibility during the financial crisis. The Journal of Finance, 72(4), 1785-1824. https://doi.org/10.1111/jofi.12505.
32. Blankespoor, E., Miller, B. P., White, H. D. (2014). The role of dissemination in market liquidity: Evidence from firms' use of Twitter. The Accounting Review, 89(1), 79-112. https://doi.org/10.2308/accr-50576.
33. Copeland, B. R., Taylor, M. S. (2004). Trade, growth, and the environment. Journal of Economic Literature, 42(1), 7-71. https://doi.org/10.1257/002205104773558047.
34. Gammerman, A., Vovk, V. (2007). Hedging predictions in machine learning: The second Computer Journal Lecture. The Computer Journal, 50(2), 151-163. https://doi.org/10.1093/comjnl/bxl065.

Appendix

Appendix A: Data Engineering and LLM Semantic Entity Fusion

A primary challenge in mapping physical climate risk to financial valuations is the structural disconnect between environmental reporting registries and financial databases. The U.S. Environmental Protection Agency's Greenhouse Gas Reporting Program (GHGRP) tracks physical emissions at the facility level, whereas the U.S. Securities and Exchange Commission's (SEC) EDGAR database aggregates financial fundamentals at the corporate parent entity level.

To bridge this gap, facility-level emissions were first aggregated to their parent company using the EPA's corporate parent linkage arrays. Because corporate naming conventions differ drastically between environmental compliance filings and SEC regulatory filings (e.g., "Wal-Mart Stores East, LP" vs. "Walmart Inc."), standard fuzzy-matching algorithms yield unacceptably high false-positive rates. To ensure deterministic accuracy, we deployed a Large Language Model (LLM) semantic fusion protocol. The LLM was prompted to execute cross-referencing based on semantic entity resolution, leveraging subsidiary hierarchies, geographic footprint data, and exact string matching where applicable. Entities that could not be mapped with absolute certainty (confidence probability < 0.85) were strictly dropped from the cross-section to prevent the introduction of synthetic noise into the predictive modeling phase.

Appendix B: Mondrian Cross-Conformal Prediction (CV+) Mathematical Formulation

To quantify the epistemic uncertainty of the XGBoost point predictions, we applied a Cross-Conformal Prediction (CV+) architecture. Standard split-conformal methods sacrifice data efficiency by requiring a separate calibration holdout. The CV+ methodology bypasses this constraint by leveraging the existing 5-Fold cross-validation structure, computing conformity scores directly from the out-of-fold predictions.

Let K be the number of folds (K=5). For each firm i, we define the non-conformity score s_i as the absolute residual between the true physical emissions y_i and the out-of-fold XGBoost prediction $\widehat{y_{-k(i)}}$:

$$s_i = \left|y_i - \widehat{y_{-k(i)}}(x_i)\right|$$

Recognizing the extreme heteroskedasticity in corporate emissions (variance scaling with asset size), an unstratified global bound would systematically violate the coverage target for heavy-tailed entities. Therefore, we applied a Mondrian condition. The covariate space was stratified into M disjoint categories (Small, Medium, and Large) based on the foundational financial constraint, Firm_Size_Log.

For a test observation X_{test} falling into Mondrian stratum m, the conformal safety net was calibrated exclusively using the non-conformity scores of the training observations within that same stratum. The dynamic marginal bound M_i required to achieve the 1-α conditional coverage (where α= 0.05 for 95% confidence) was calculated as the empirical quantile of the sorted non-conformity scores:

$$\widehat{q_m} = \text{Quantile}\left(\{s_j : j \in \text{Stratum } m\}, \frac{\lceil (N_m + 1)(1 - \alpha) \rceil}{N_m}\right)$$

This mathematical stratification guarantees that the resulting lower bound—the foundation of the Conformal-Weighted Continuous Divergence (CWCD) score utilized in the Stage 2 econometrics—is dynamically calibrated to the exact financial constraints of the underlying asset class.

Appendix C: Extended Econometric Robustness

The primary causal inference specifications detailed in Section 4.6 (Tables 8 and 9) utilized ordinary least squares (OLS) regressions parameterized with high-dimensional Industry Fixed Effects and heteroskedasticity-robust standard errors (HC3). Continuous variables were Winsorized at the 1% and 99% thresholds to neutralize outliers. To ensure the algorithmic divergence penalty (CWCD) was not a fragile artifact of these specific econometric parameters, extended robustness checks were executed.

Alternative Standard Error Clustering: The primary models were re-estimated utilizing firm-level clustering rather than simple HC3 robust errors to account for potential unobservable intra-firm residual correlation. The coefficient on ACD_Score_T remained deeply negative and highly significant ($p < 0.01$) across both the Tobin's Q and ROA specifications.

Aggressive Winsorization Thresholds: To ensure the results were not driven by the long tails of the remaining distribution, the models were re-estimated with an aggressive 5% and 95% Winsorization threshold. The economic magnitude of the valuation penalty (Tobin's Q) and operational decay (ROA) persisted, confirming that the market discipline effect is a structural reality across the entire distribution, rather than a phenomenon isolated to extreme outliers.