Corporate Omissions: Correcting the Bias in Carbon Reporting

Giulia Crippa[1], Florian Berg[2], and Roberto Rigobon[2]

[1]Operations Research and Financial Engineering, Princeton University
[2]MIT Sloan School of Management

February 26, 2026

Abstract

Corporate carbon disclosure is central to asset pricing and sustainable finance, yet firm-level emissions data remain incomplete and selectively reported. We model disclosure as a selection process in which reporting depends on latent emissions and implement a missing-not-at-random (MNAR) correction that recovers the full emissions distribution rather than relying on conventional missing-at-random imputations. Using U.S. public firms data from 2012–2023, we show that higher-emitting firms are systematically less likely to disclose. Correcting for this selection increases average Scope 1 emissions by 15–19%, with substantial mass added to the upper tail, implying that standard approaches materially understate aggregate emissions and associated carbon damages. Importantly, imputing missing emissions with MNAR-adjusted values leaves return–emissions relationships largely unchanged, indicating that the correction does not induce spurious pricing effects. Our results highlight selective disclosure as a first-order source of bias in corporate carbon data with direct implications for empirical asset pricing and climate policy.

1 Introduction

Accurate measurement of corporate carbon emissions is central to climate policy, sustainable finance, and corporate accountability. As governments, investors, and stakeholders increasingly demand transparency in environmental performance, Scope 1 emissions – direct greenhouse gas emissions from sources owned or controlled by firms – have emerged as a foundational metric for assessing corporate contributions to global carbon footprints. Recent regulatory initiatives, including the European Union's Corporate Sustainability Reporting Directive (CSRD) and the U.S. Securities and Exchange Commission's proposed climate disclosure rules, aim to mandate and standardize emissions reporting. However, implementation remains incomplete: the SEC's rule is subject to ongoing judicial review, while key CSRD obligations were formally delayed in April 2025. As a result, incomplete and uneven emissions disclosure continues to characterize corporate carbon data.

Despite growing standardization efforts, corporate carbon accounting remains fragmented in both coverage and quality. Firms face heterogeneous incentives to disclose emissions, shaped by regulatory exposure, reputational concerns, and potential financial consequences (Christensen et al., 2021). In particular, high-emitting firms may strategically withhold emissions data to avoid scrutiny or adverse stakeholder reactions. Even among disclosers, verification practices vary substantially. Recent evidence shows that firms obtaining third-party assurance report significantly higher emissions than non-assured firms, suggesting that unaudited disclosures may systematically understate true emissions levels (Berg et al., 2024). These patterns raise concerns about the reliability of both self-reported and vendor-estimated emissions data.

A growing literature studies corporate carbon emissions and their role in financial markets, sustainability reporting, and climate policy. Prior work documents substantial heterogeneity in emissions across firms and sectors, examines the determinants of disclosure, and evaluates whether carbon intensity is priced in equity markets (Aswani et al., 2024; Bolton & Kacperczyk, 2021; Pastor et al., 2025). Related research highlights pervasive measurement challenges in environmental data, including inconsistent reporting standards, limited verification, and reliance on vendor estimates cite. To address missing emissions data, most empirical studies rely on imputation strategies that implicitly assume emissions are missing at random, conditional on observable firm characteristics (Assael et al., 2023; Duan & Pelger, 2025). While convenient, this assumption is difficult to reconcile with voluntary disclosure environments, where reporting decisions are endogenous and potentially strategic (Dye, 1985; Verrecchia, 1983). As a result, existing approaches may systematically understate emissions for non-disclosing firms and distort aggregate assessments of corporate carbon footprints.

Most empirical work addressing missing emissions relies on imputation methods that assume data are missing completely at random (MCAR) or missing at random (MAR), implying that the probability of missingness is independent of the unobserved emissions themselves. In the context of voluntary disclosure, this assumption is difficult to justify. Firms' incentives to disclose are plausibly linked to the magnitude of their emissions, consistent with a missing-not-at-random (MNAR) mechanism. Ignoring this dependence can lead to systematic underestimation of aggregate emissions, distorted sectoral allocations, and misleading assessments of environmental impact. These concerns are quantitatively important: in 2022, the U.S. Environmental Protection Agency reported total national greenhouse gas emissions of 6,343 million metric tons of CO_2 equivalent, yet the contribution of publicly listed firms – particularly those that do not disclose – remains poorly quantified due to data gaps.

This paper makes three contributions. First, we develop a principled framework for imputing corporate Scope 1 emissions under non-ignorable missingness, explicitly modeling disclosure decisions as a function of latent emissions rather than assuming ignorable reporting. Our approach departs from standard MAR-based imputations used in the sustainability and asset-pricing literatures by allowing for strategic non-disclosure and correcting the resulting bias through an exponential tilting mechanism. Second, we apply this framework to a comprehensive panel of U.S. public companies from 2012 to 2023, producing revised estimates of the level and distribution of Scope 1 emissions that materially alter aggregate emissions and their sectoral composition. Third, we provide external validation of the framework using firms that transition from non-disclosure to disclosure and a diagnostic return-based exercise, showing that MNAR correction reshapes emissions distributions without inducing spurious correlations in downstream empirical applications.

Empirically, we combine emissions data from Trucost Environmental with firm-level financial information from Compustat, covering approximately 8,300 U.S. public companies per year. Our methodology proceeds in five steps. We first estimate an outcome model for emissions using the subsample of disclosing firms. We then impute missing emissions and estimate a disclosure response mechanism that allows reporting probabilities to depend on firm characteristics and latent emissions. Identification is achieved using aggregate time variation, including calendar time and U.S. economic activity, under the assumption that disclosure decisions are relatively sticky once firm size, sector, and reporting regimes are controlled for. We implement both fully parametric and semi-parametric specifications of the response mechanism to accommodate potential nonlinearities. Finally, we apply exponential tilting to adjust the imputed emissions distribution for selective disclosure.

The results reveal pronounced and systematic patterns in emissions disclosure. Between 2012 and 2023, the share of firms directly disclosing Scope 1 emissions increased from roughly 4% to 17%, while the fraction of firms with missing emissions data declined but remained substantial. Disclosure rates vary sharply across sectors and firm size: energy-intensive sectors such as Energy and Industrials disclose more frequently, whereas Financials and Real Estate exhibit persistently low reporting rates, particularly among smaller firms. Our MNAR-based imputations indicate that non-disclosing firms tend to have higher emissions, leading to a 15–19% increase in average Scope 1 emissions relative to MAR-based imputations. Using firms that transition from non-disclosure to disclosure, we show that MNAR-adjusted estimates reduce systematic underestimation in the upper tail of the emissions distribution, where the most carbon-intensive firms are concentrated.

As an application of the revised emissions estimates, we quantify corporate carbon damages following the framework of Greenstone et al. (2023). Using alternative social cost of carbon assumptions, we document substantial aggregate damages attributable to U.S. public companies and a marked concentration of damages in a small set of high-emission sectors. While these estimates are not directly comparable to economy-wide inventories, they underscore the quantitative importance of accounting for selective disclosure when assessing the societal costs of corporate emissions.

Finally, we use a standard return regression as a diagnostic exercise to assess whether MNAR-based imputations distort empirical relationships in which emissions are commonly used as regressors. Replacing observed emissions with MNAR-imputed values leaves the return–emissions relationship essentially unchanged, indicating that the imputation procedure preserves existing empirical relationships. In contrast, the decision to disclose emissions carries a robust return premium, consistent with classic theories of voluntary disclosure (Dye, 1985; Grossman, 1981; Milgrom, 1981; Verrecchia, 1983).

By modeling emissions disclosure as a non-ignorable process, this paper challenges the prevailing reliance on ignorable missingness assumptions in corporate carbon accounting. Our results highlight the dominant role of selective disclosure in shaping observed emissions data and underscore the importance of targeted policies to improve reporting among high-emission firms. More broadly, the framework developed here provides a flexible tool for addressing missing data in environmental measurement settings where disclosure incentives are intrinsic to the outcome of interest.

The remainder of the paper proceeds as follows. Section 2 discusses the imputation of corporate emissions. Section 3 presents the empirical framework, detailing the modeling of disclosure as a missing-not-at-random process and the imputation strategy used to correct for endogenous non-reporting. Section 4 introduces the data, while Section 5 describes the empirical methodology. Section 6 reports the main results, and Section 7 discusses the implications for aggregate emissions and corporate carbon damages. Section 8 provides robustness checks for the return–emissions relationship. Section 9 concludes.

2 On the Imputation of Corporate Emissions

Imputing corporate carbon emissions is not merely a statistical exercise. Unlike financial data, emissions disclosure is largely voluntary, and therefore informative about firms' underlying environmental performance (Shi et al., 2025). This section focuses on the failure of standard approaches to emissions imputation when reporting incentives are endogenous, and explains why these methods tend to underestimate aggregate pollution.

2.1 The implicit MAR assumption in the literature

Most empirical studies relying on firm-level emissions data face substantial missingness, as corporate carbon disclosures remain largely voluntary and incomplete. Researchers often address this challenge through simple and traditional imputation techniques. These include imputing emissions using sector-year averages or medians scaled by firm size (Bolton & Kacperczyk, 2021; Pastor et al., 2025), relying on vendor-provided estimates constructed from benchmarking and engineering models (LSEG, 2024; S&P Global, 2025), or applying machine-learning algorithms trained on firms that disclose emissions (Assael et al., 2023; Nguyen et al., 2021).

However, these approaches share a common and often unstated assumption: conditional on observable characteristics such as sector or size, firms that do not report emissions are treated as statistically comparable to those that do. According to Rubin (1987)'s classification of missingness, this corresponds to the ignorable or missing-at-random (MAR) assumption. Under MAR, the distribution of emissions among non-reporting firms can be fully recovered from the observed sample without modeling the disclosure decision itself.

The ignorability assumption is rarely discussed explicitly in the sustainability or finance literature. As long as non-reporting is unrelated to missing emissions—after conditioning on observables—such methods are internally consistent. The key question, therefore, is whether this assumption is reasonable in the context of corporate emissions reporting.

2.2 Drivers of corporate emissions disclosure

Corporate emissions disclosure is shaped by a combination of regulatory, economic, and reputational incentives. Unlike financial reporting, carbon disclosure remains largely voluntary in many jurisdictions, and firms differ substantially in both their ability and willingness to report emissions. Understanding these drivers is essential for assessing whether missing emissions can plausibly be treated as random.

Regulatory exposure is a primary determinant of disclosure. Firms subject to environmental disclosure regulation are significantly more likely to report emissions, while firms facing limited regulatory oversight are more likely to remain silent (Christensen et al., 2021; Greenstone et al., 2023). Anticipation of future regulation can also induce preemptive disclosure (Downar et al., 2021). Importantly, the introduction of mandatory disclosure regimes is often followed by sharp increases in reported emissions, suggesting that prior non-disclosure reflected reporting incentives rather than low pollution levels.

Reputational and stakeholder pressures further shape disclosure decisions. Firms with high public visibility, consumer-facing business models, or substantial institutional ownership face stronger incentives to appear environmentally responsible and are therefore more likely to disclose emissions (Dhaliwal et al., 2011; Lyon & Montgomery, 2015). Investor engagement and pressure from non-governmental organizations have also been shown to increase both the likelihood and scope of environmental disclosure (Dyck et al., 2019).

Disclosure additionally entails direct and indirect costs. Measuring emissions requires internal data collection, coordination across production units, and often third-party verification. These costs increase with operational complexity and emissions intensity, particularly for firms with diversified operations or complex supply chains (Halkos & Nomikos, 2021). Consistent with this view, firms obtaining third-party assurance report significantly higher emissions than otherwise comparable firms, indicating that unaudited disclosures tend to understate pollution levels (Berg et al., 2024).

Taken together, these findings imply that emissions disclosure is not a passive reporting outcome, but a strategic decision that varies systematically with firm characteristics, regulatory exposure, and pollution intensity. Missing emissions data therefore reflect underlying economic incentives rather than random reporting gaps.

Treating emissions as missing at random has direct and non-trivial consequences. The next subsection examines how MAR-based imputation leads to systematic underestimation of emissions.

2.3 Strategic non-disclosure and right-tail selection

When disclosure is endogenous, there are strong economic reasons to doubt that CO_2 emissions data are randomly missing (Christensen et al., 2021). Although disclosure can mitigate information asymmetries between firms and stakeholders, it also entails direct and indirect costs that may offset its benefits. Reporting high emissions can expose firms to reputational damage, regulatory scrutiny, and litigation risk. As a result, disclosure is a strategic choice rather than a passive reporting exercise.

Classic disclosure theory emphasizes that, in voluntary regimes, non-disclosure is informative. When agents can choose whether to reveal information, silence is often interpreted as unfavorable news (Milgrom, 1981; Verrecchia, 1983). In the context of emissions reporting, this mechanism implies that firms with higher emissions face stronger incentives to withhold disclosure.

Empirical evidence from the sustainability literature supports this interpretation. Disclosure rates are negatively related to pollution intensity across firms and sectors (Halkos & Nomikos, 2021). Firms obtaining third-party assurance report significantly higher emissions than otherwise comparable firms, suggesting systematic understatement in unaudited disclosures (Berg et al., 2024). Moreover, mandatory disclosure regimes are frequently associated with upward revisions in reported emissions (Greenstone et al., 2023). Together, these findings indicate that non-disclosure is not random, but correlated with the level of unobserved emissions.

This strategic mechanism generates a consequential form of selection: the observed emissions distribution is truncated from above. High emitters are under-represented among reporting firms, and standard imputation methods therefore understate true emissions.

Imputations based on sector-year medians necessarily shrink predicted emissions toward the center of the observed distribution. When the right tail is selectively missing, these averages lie below the true conditional mean for non-reporting firms. This problem is amplified by the extreme right-skewness of emissions distributions, where a small fraction of firms accounts for a disproportionate share of aggregate pollution (Fankhauser, 1994). As a result, imputations drawn from the observed sample systematically miss upper-tail mass. This concentration of emissions among large polluters is reflected in investor and policy initiatives that explicitly target the largest corporate emitters.

Importantly, underestimation is not the result of limitations in imputation methodologies. It follows directly from ignoring the disclosure decision itself. When disclosure decisions are not accounted for, the resulting bias is not model-specific: truncation necessarily leads to underestimation of both firm-level and aggregate emissions. Correcting this bias therefore requires modeling the disclosure decision explicitly, rather than relying on increasingly flexible imputations applied to a selectively observed sample. The next section develops such a framework.

3 An MNAR Framework for Emissions Disclosure

This section introduces the formal framework used to model corporate emissions disclosure when missingness is non-ignorable. Motivated by the evidence in Section 2 that emissions disclosure is endogenous and correlated with underlying pollution levels, we treat missing emissions as arising from a selection process that depends on unobserved outcomes. We outline the disclosure mechanism, state the assumptions required for identification, and introduce the exponential tilting representation that underpins our empirical approach.

3.1 Disclosure as a response mechanism

Let y_i denote firm i's true Scope 1 CO_2 emissions, and let $\mathbf{x}_i$ be a vector of observed firm characteristics. Emissions are observed only when firms disclose them. Let $\delta_i \in \{0, 1\}$ be an indicator variable of whether y is observed. In our setup, $\mathbf{x}_i$'s are always available, while y_i's are observable only when $\delta_i = 1$; otherwise, when $\delta_i = 0$, the corresponding

y_i is missing. The disclosure decision is governed by a response mechanism

$$\mathbb{P}(\delta_i = 1 \mid \mathbf{x}_i, y_i), \tag{1}$$

which captures the probability that emissions are observed conditional on firm characteristics and the emissions level itself.

When disclosure depends on y_i after conditioning on observables, missingness is non-ignorable (MNAR). In this case, the distribution of observed emissions differs systematically from the distribution of unobserved emissions, and standard imputation methods that ignore the disclosure process are generally biased. As explained in Section 2, this setting is natural for corporate emissions data, where firms with higher emissions face stronger incentives to withhold disclosure.

3.2 Identification under MNAR

Statistically, two are the main challenges associated with MNAR missingness: (i) no test exists to confirm MNAR since the missing values are, by definition, unobserved (Little, 1988); (ii) the joint distribution of $(\mathbf{x}, y, \delta)$ is typically non-identifiable without additional assumptions. In practice, the following two conditions are sufficient for identification.

First, we assume a parametric specification for the response mechanism, indexed by a finite-dimensional parameter vector γ:

$$\mathbb{P}(\delta_i = 1 \mid \mathbf{x}_i, y_i) = \pi(\delta_i = 1 \mid \mathbf{x}_i, y_i; \gamma), \tag{2}$$

where $\pi(\cdot)$ is a known link function. This provides the minimal structure required to separate the outcome process from the disclosure process. Moreover, a parametric response model makes explicit how disclosure incentives vary with observable firms characteristics and the level of emissions itself.

Second, as shown by Wang et al. (2014), we assume the existence of excluded covariates that affect emissions but do not directly influence disclosure once core firm characteristics are controlled for. Specifically, let $\mathbf{x}_i = (\mathbf{x}_{1i}, \mathbf{x}_{2i})$, where $\mathbf{x}_{1i}$ is contained in $\mathbf{x}_{2i}$. Then, the additional variables in $\mathbf{x}_{2i}$ shift emissions but are conditionally independent of disclosure:

$$\mathbb{P}(\delta_i = 1 \mid \mathbf{x}_i, y_i) = \mathbb{P}(\delta_i = 1 \mid \mathbf{x}_{1i}, y_i). \tag{3}$$

This exclusion restriction provides variation that separates the outcome process from the disclosure process and allows the joint distribution to be identified. The empirical plausibility of this assumption is discussed in Section ??, and its implications are assessed through sensitivity analysis.

3.3 Exponential tilting

Given identification, the key implication of MNAR disclosure is that the distribution of unobserved emissions differs from that of observed emissions in a systematic way. Let

$$f_1(y \mid \mathbf{x}) \equiv f(y \mid \mathbf{x}, \delta = 1) \quad \text{and} \quad f_0(y \mid \mathbf{x}) \equiv f(y \mid \mathbf{x}, \delta = 0)$$

denote the conditional densities of observed and unobserved emissions, respectively. By Bayes' rule, these densities are related through the odds of non-disclosure,

$$f_0(y \mid \mathbf{x}) = f_1(y \mid \mathbf{x}) \cdot \frac{O(\mathbf{x}, y)}{\mathbb{E}[O(\mathbf{x}, y) \mid \mathbf{x}, \delta = 1]}, \tag{4}$$

where $O(\mathbf{x}, y) = \mathbb{P}(\delta = 0 \mid \mathbf{x}, y)/\mathbb{P}(\delta = 1 \mid \mathbf{x}, y)$.

We adopt an exponential response model of the form

$$\mathbb{P}(\delta_i = 1 \mid \mathbf{x}_i, y_i) = \frac{\exp\{f(\mathbf{x}_{1i}) + g(y_i)\}}{1 + \exp\{f(\mathbf{x}_{1i}) + g(y_i)\}}, \tag{5}$$

where $f(\cdot)$ and $g(\cdot)$ capture the effects of observables and emissions on disclosure. Under this specification, the relationship between observed and unobserved emissions simplifies to an exponential tilting of the observed distribution:

$$f_0(y \mid \mathbf{x}) = f_1(y \mid \mathbf{x}) \cdot \frac{\exp\{-g(y)\}}{\mathbb{E}[\exp\{-g(y)\} \mid \mathbf{x}, \delta = 1]}. \tag{6}$$

This representation makes the nature of the selection problem transparent. When higher emissions reduce the probability of disclosure, $g(y)$ is increasing in y, and the unobserved distribution places relatively more mass in the right tail. Ignoring this tilting mechanism therefore leads to systematic understatement of emissions, particularly at the aggregate level. The exponential form provides a tractable and economically interpretable way to correct this bias.

In survey research, imputation methods vary based on assumptions on $f(\cdot)$ and $g(\cdot)$ in (5). Fully parametric approaches assume specific forms for both the outcome model $f(\cdot)$ and the response model $g(\cdot)$, requiring both to be correctly specified (Kott, 2006;

Qin et al., 2008). To mitigate risks of misspecification, recent advances leverage semi-parametric methods, which offer flexibility in modeling non-ignorable missingness without strict distributional assumptions on the outcome (Yang & Kim, 2017).

In this paper, we consider two specifications of the disclosure mechanism: a fully parametric model and a semi-parametric alternative. Comparing these approaches allows us to assess the importance of nonlinear disclosure responses by contrasting linear specifications with flexible generalized additive models (GAMs).

The next sections describe the empirical implementation of this framework. Section 4 introduces the firm-level emissions data, and Section 5 presents the empirical specification, including the outcome model, the disclosure mechanism, and the exponential tilting adjustment.

4 Data

This section describes the data sources, sample construction, and the definition of observed and missing emissions used in the analysis. We focus on Scope 1 CO_2 emissions for U.S. publicly listed firms and restrict attention to directly reported emissions in order to isolate disclosure behavior.

4.1 Sample

We combine firm-level emissions data from Trucost Environmental with financial information from Compustat. Trucost provides detailed coverage of corporate greenhouse gas emissions, while Compustat supplies balance sheet and income statement variables used to characterize firm size and industry affiliation. Our sample consists of U.S. public companies observed annually from 2012 to 2023. After merging the two datasets, the final sample includes approximately 8,300 firms per year, for a total of 102,337 firm-year observations.

We focus exclusively on Scope 1 emissions, defined as direct emissions from sources owned or controlled by the firm under the Greenhouse Gas Protocol. Scope 2 and Scope 3 emissions are excluded due to substantial measurement challenges, heterogeneity in reporting standards, and the risk of double counting (Cheema-Fox et al., 2021). Restricting attention to Scope 1 emissions allows for a more precise assessment of corporate disclosure behavior and limits contamination from upstream or downstream reporting practices.

Table A1 provides annual summary statistics for the sample. The number of covered firms declines gradually over the sample period, consistent with broader trends in U.S. equity markets, while total market capitalization increases substantially, particularly after 2020. Aggregate Scope 1 emissions remain relatively stable over time, whereas average emissions intensity declines markedly, reflecting a combination of changes in firm composition, scale effects, and improvements in emissions efficiency.

Observed versus missing emissions A key feature of the Trucost dataset is that emissions values originate from multiple sources. In addition to firm-reported emissions – sourced from sustainability reports, CDP disclosures, or direct company communication – Trucost provides model-based estimates constructed using industry benchmarks such as the and engineering assumptions. Because these estimates do not reflect firms' disclosure choices, we treat them as missing for the purposes of this study.

Specifically, we classify emissions as *observed* only when they are labeled as *exact* or *derived* by Trucost. Emissions values classified as *estimated* are treated as missing. This distinction allows us to isolate firms that actively disclose emissions and to study non-disclosure as a selection process rather than as a measurement error problem. Table A2 reports the detailed provenance of Scope 1 emissions data in the Trucost dataset for the year 2023, together with our corresponding classification.

In addition, we classify as missing the emissions of firms that appear in Compustat but are not covered by Trucost Environmental. This choice expands the analysis to a broader cross-section of U.S. public firms and ensures meaningful representation of the overall economy. In contrast, related studies typically restrict attention to firms with emissions coverage or rely on vendor-provided estimates, thereby excluding a substantial fraction of non-reporting firms (Chen et al., 2025; Duan & Pelger, 2025; Olesiewicz et al., 2021). Appendix Table A3 reports summary statistics separately for firms with observed emissions and firms with missing emissions.

4.2 Patterns in emissions disclosure

To better model disclosure, it is of paramount importance to analyze fundamental patterns about what firms disclose, when disclosure occurs, and how coverage evolves across firms characteristics. Understanding these mechanisms provides essential foundations for subsequent modeling assumptions, and clarifies the extent to which ob-

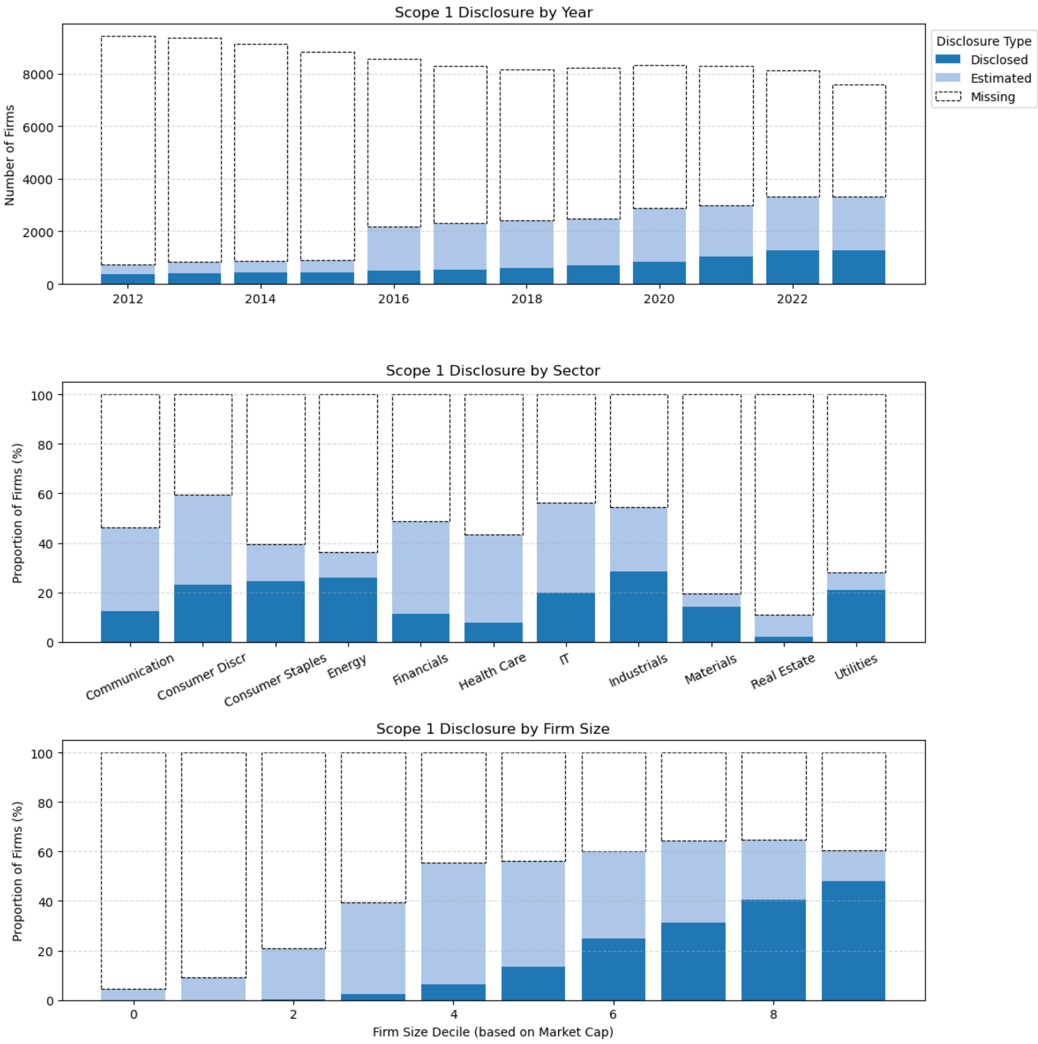

Figure 1: Trucost Scope 1 Disclosure Trend with respect to year (top graph), sector (middle graph) and firm size (bottom graph).

served emissions reflect reporting behavior rather than underlying production technology.

Figure 1 summarizes patterns in Scope 1 emissions disclosure over time, across sectors, and by firm size. The top panel documents a steady increase in the fraction of firms that disclose emissions directly, rising from approximately 4% in 2012 to 17% in 2023. Over the same period, the share of firms for which Trucost provides estimated emissions expands more sharply, particularly after 2016, reflecting both expanded data coverage and heightened attention to climate reporting following the Paris Agreement. Despite these trends, a substantial fraction of firms continue to have missing emissions data throughout the sample period.

The middle and bottom panels focus on cross-sectional patterns in 2023. The middle panel reports disclosure rates by sector and reveals pronounced heterogeneity across industries. Industrials, Energy, and Consumer Staples exhibit the highest disclosure rates, while Financials, Health Care, and Real Estate display substantially lower coverage. These differences are consistent with sectoral variations in emissions intensity, regulatory exposure, and stakeholder scrutiny. Notably, energy-intensive sectors appear among the leading discloser of Scope 1 emissions.

The bottom panel reports disclosure rates by firm size deciles, measured by market capitalization. Larger firms are substantially more likely to disclose emissions or to be covered by Trucost estimates, whereas disclosure among smaller firms remains limited. In the bottom three size deciles, fewer than 20% of firms have any emissions data available, and direct disclosure is rare. In contrast, firms in the top deciles exhibit disclosure rates exceeding 60%, with direct reporting accounting for a large share.

Together, these patterns highlight systematic variation in emissions disclosure across time, sectors, and firm size, underscoring the importance of accounting for selective disclosure in empirical analyses. In the next section, we describe the empirical framework used to address this selection when imputing missing emissions.

5 Empirical Framework

This section describes the empirical framework used to impute missing Scope 1 emissions under non-ignorable disclosure. Guided by the identification strategy in Section 3.2, the approach proceeds in five steps: (i) estimation of an outcome model using observed emissions, (ii) imputation of missing emissions under a baseline MAR assumption, (iii) estimation of the disclosure (response) mechanism, (iv) correction for selection via exponential tilting, and (v) sensitivity analysis. Figure 2 provides an overview of the procedure.

5.1 Outcome model

Let y_i denote Scope 1 CO_2 emissions for firm i, and let $\mathbf{x}_i$ denote a vector of observable covariates. The first step of the procedure estimates the conditional distribution of emissions among disclosing firms. We specify a linear outcome model,

$$y_i = \eta_0 + \mathbf{x}_i^\top \eta + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \tag{7}$$

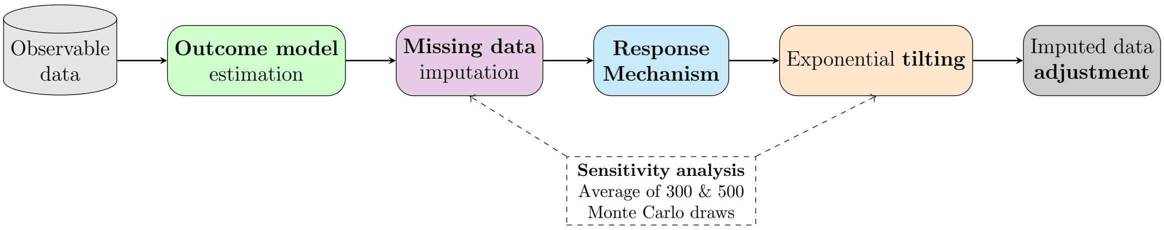

Figure 2: Workflow of the missing data imputation methodology. The procedure consists of five main steps: (1) outcome modeling based on observable data; (2) imputation of missing values; (3) estimation of the response mechanism; (4) application of exponential tilting; and (5) adjustment of the imputed values. Additionally, a sensitivity analysis is conducted using Monte Carlo simulations with 300 and 500 random draws, respectively.

where $\mathbf{x}_i$ includes firm size (proxied by total assets), sector fixed effects (based on the Global Industry Classification Standard, GICS), time (measured in years), and level of the aggregate economic activity (U.S. GDP). We use total assets rather than market capitalization to proxy for production scale and to avoid valuation-driven volatility. GICS, with 11 defined sectors, facilitates global comparability to support future extensions beyond US firms. Time and U.S. GDP are included to capture macroeconomic and reporting cycles.

A useful way to think about this first step is as recovering the production-side signal in emissions that is observable among firms that choose to disclose. Conditional on disclosure, Scope 1 emissions are largely driven by scale, industry technology, and aggregate economic conditions. Equation (7) formalizes this intuition by mapping emissions to observables that proxy for these underlying determinants, while treating residual variation as idiosyncratic noise.

Importantly, Equation (7) is estimated on the subsample of disclosing firms, to obtain the consistent estimator $\hat{\eta} = (\hat{\eta}_0, \hat{\eta}_1)$ that characterizes the conditional distribution of emissions $f_1(y|\mathbf{x}, \hat{\eta})$. This step does not attempt to model disclosure behavior itself; rather, it isolates the conditional emissions process that would prevail absent selection. Under MAR, this model would suffice to impute missing values.

In the presence of non-ignorable disclosure, this step serves as a baseline to generate inital imputations for non-reporting firms. Specifically, we draw from the predictive distribution

$$\hat{y}_i \mid \mathbf{x}_i \sim \mathcal{N}\left(\hat{\eta}_0 + \mathbf{x}_i^\top \hat{\eta}, \hat{\sigma}^2\right). \tag{8}$$

This step generates counterfactual emissions under the assumption that non-reporting firms follow the same conditional distribution as reporting firms, i.e., $f_1(y|\mathbf{x}, \hat{\eta}) = f_0(y|\mathbf{x}, \hat{\eta})$. As discussed in Section 2, this assumption is generally violated when disclosure is endogenous. The subsequent steps adjust these imputations to account for selection.

5.2 Disclosure mechanism

As discussed in Section 3.2, identification of the joint distribution $(\mathbf{x}, y, \delta)$, where δ is an indicator variable for disclosure, requires the existence of at least one instrumental variable $\mathbf{x}_2$ such that $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$ and $\mathbf{x}_2$ affect emissions but not disclosure after $\mathbf{x}_1$ is considered.

In this empirical implementation, time and U.S. GDP serve as excluded shifters of emissions. While these variables may correlate with secular changes in disclosure norms, we condition on firm size and sector to absorb persistent differences in disclosure incentives. Conditional on these controls, disclosure decisions are assumed to be relatively sticky, reflecting fixed reporting costs and institutional infrastructure, so that residual time variation operates primarily through emissions rather than disclosure. The plausibility of this restriction is assessed through sensitivity analysis.

Specifically, we model the probability of emissions disclosure using a response mechanism that depends on firm characteristics and emissions:

$$\mathbb{P}(\delta_i = 1 \mid \mathbf{x}_{1i}, y_i) = \frac{\exp\{f(\mathbf{x}_{1i}) + g(y_i)\}}{1 + \exp\{f(\mathbf{x}_{1i}) + g(y_i)\}}, \tag{9}$$

where δ_i is the disclosure indicator, $\mathbf{x}_{1i}$ includes firm size and sector, and $g(y_i)$ captures the dependence of disclosure on emissions. We consider two specifications of the response mechanism. A Fully Parametric (FP) case, where both $f(\mathbf{x}_{1i})$ and $g(y_i)$ are linear functions. A semi-parametric approach, where $f(\cdot)$ and $g(\cdot)$ are estimated flexibly using generalized additive models (GAMs), such that penalized splines allow for nonlinear disclosure responses.

5.3 Exponential tilting adjustment

Given estimates of the response mechanism, we correct for selection using exponential tilting. Under the logistic specification in Equation (9), the distribution of unobserved emissions differs from the observed distribution by a tilting factor proportional to $\exp\{-g(y)\}$. We therefore reweight imputed emissions by

$$w_i \propto \exp\{-\hat{g}(y_i)\}, \tag{10}$$

with tilting factors derived from the estimated parameters of the response function.

Intuitively, firms with emissions levels associated with a lower probability of disclosure receive greater weight in the adjusted distribution. This procedure restores right-tail mass that is systematically under-represented among reporting firms and yields selection-corrected estimates of firm-level and aggregate emissions.

5.4 Sensitivity analysis

Because MNAR assumptions cannot be tested directly, we evaluate robustness through Monte Carlo sensitivity analysis. In fact, even when the instrument is believed to satisfy the exclusion restriction, potential model misspecification and unobserved heterogeneity may still bias inferences about imputed data (Little & Rubin, 2019). To quantify this uncertainty, we repeatedly draw imputations from Equation (8), re-estimate the response mechanism, and apply exponential tilting. We conduct this procedure using 300 and 500 random draws to construct empirical distributions of coefficient estimates, assess their stability, and quantify the variation of first and second moments of the imputed emissions (Gohil, 2020).

6 Results

This section presents the main empirical findings on CO_2 emissions disclosure among U.S. public companies. Building on the outcome model, response mechanism, and exponential tilting framework developed in the previous section, we quantify the implications of selective disclosure for inferred emissions and assess the implications of MNAR-based bias correction for inferred emissions distributions.

We organize the results as follows. First, we report estimates from the outcome model of Scope 1 CO_2 emissions, which provides the baseline conditional distribution used to impute missing values. Second, we examine the estimated disclosure response mechanism, comparing fully parametric and semi-parametric specifications and highlighting their implications for selection. These estimates inform the construction of tilting factors used to correct for non-random disclosure. Finally, we evaluate how exponential tilting reshapes the imputed emissions distribution relative to conventional MAR-based imputations, with particular attention to changes in the level and upper tail of inferred emissions.

6.1 Outcome model

The first step of our methodology involves imputing missing data using the respondent portion of the sample. To ensure a more stable relationship between emissions and firm characteristics, both scope 1 emissions and total assets have been log-transformed. Table A4 presents the estimated coefficients from our model.

All covariates, with the exception of a few year indicators, are statistically significant at the 5% level. The model explains a substantial portion of the variance in reported CO_2 emissions, with an adjusted R^2 of 0.671.

Firm size, proxied by log assets, shows a strong positive association with Scope 1 emissions. The coefficient estimate of 0.997 suggests that a 1% increase in firm size is associated with approximately a 0.997% increase in emissions, scaling almost linearly. Conversely, higher aggregate economic activity (denoted as GDPC1) is modestly associated with increased emissions at the firm level, with a positive and statistically significant coefficient of 0.0497.

Sectoral differences in emissions are substantial. Communication Services serves as the baseline category in the regression model. Relative to this group, Utilities, Materials, and Energy firms exhibit the largest positive differentials in log emissions, with coefficients of 5.42, 4.66, and 4.90, respectively. These values suggest that, holding other factors constant, firms in these sectors emit substantially more than those in the baseline group. In contrast, Financials report significantly lower emissions, with a coefficient of -2.75, while IT and Health Care sectors show emissions levels comparable to the baseline, with modest positive coefficients of 0.49 and 0.66, respectively. This pattern is consistent with the broad distinction between more emission-intensive (*brown*) and less emission-intensive (*green*) sectors, reflecting structural differences in business operations and energy use.

from the regression model indicate a clear and statistically significant downward trend in reported Scope 1 emissions conditional on firm characteristics over the past decade. Using 2012 as the reference year, the coefficients for subsequent years become increasingly negative, with effects growing in magnitude and statistical significance over time. While early-year coefficients (2013–2016) are small and not statistically different from zero, the decline becomes significant starting in 2018. By 2023, the coefficient reaches -0.82, implying that, all else equal, reported emissions in 2023 were approximately 56% lower than in 2012.

This sustained decline likely reflects a combination of structural changes, including tightening climate regulation, increased stakeholder scrutiny, and improvements in emissions measurement and reporting. The accelerating pace of change after 2018 may also coincide with major global climate initiatives and improved emissions tracking mechanisms. These time effects offer strong evidence that the landscape of corporate emissions disclosure has evolved materially over the past decade.

Following our methodology, the estimated outcome model is used to generate plausible values for missing emissions data by randomly drawing from the predictive distribution in Equation (7). This forms the first step of the imputation process, to obtain a complete dataset that will be utilized to model the response mechanism.

6.2 Response Mechanism

After imputing scope 1 emissions, we proceed to estimate the response propensity for each company, leveraging the complete covariates $\mathbf{x}_1$ and the dependent variable y. Specifically, the primary objective is to calculate the probability that a company reports its scope 1 emissions based on its size, sector, and emissions value. Specifically, the objective is to estimate the probability that a company reports its Scope 1 emissions as a function of firm characteristics and emissions levels. The inclusion of emissions as a predictor reflects the maintained hypothesis that disclosure decisions may depend on the latent emissions outcome, consistent with non-ignorable missingness. Importantly, this specification does not test the MNAR assumption, but parameterizes a class of selection mechanisms that allow disclosure to vary systematically with emissions.

Our methodology incorporates an exponential tilting technique to adjust for potential biases in the outcome model predictions, which may underestimate emissions for non-disclosing companies. This adjustment is based on the hypothesis that companies are less likely to disclose emissions when their values are high. According to Section 5 we employ a fully parametric approach versus a semi parametric one to estimate the response mechanism coefficients as in Equation (9). Below we discuss our results.

Fully Parametric Figure 3 presents the distribution of the estimated coefficients for 300 Monte Carlo random draws, as specified in Equation (7) (see Table A5 for detailed figures and results for 500 draws). These results align with the existing literature on sustainability reporting, revealing that firm size and emissions are associated with higher and lower disclosure propensities, respectively. Specifically, larger firms tend to report their CO_2 scope 1 emissions more frequently, whereas firms with higher emissions are less likely to disclose their environmental impact.

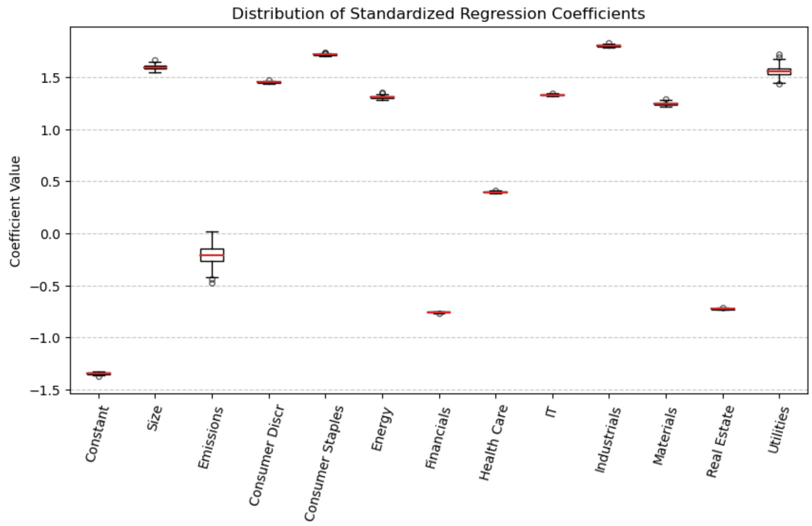

Figure 3: Distribution of regression coefficients of the response mechanism model, based on 300 random draws, for the fully-parametric model. The boxplot displays the variability in the estimated coefficients for the constant term, firm size, emissions, and sector indicator. Both firm size and emission coefficients are scaled by the standard deviation of the correspondent variable. The red line represents the mean coefficient across random iterations, while the whiskers and outliers illustrate the range of variability.

Since log-odds are not directly interpretable, we translate them into probabilities to facilitate our analysis. The results of our model indicate that the average probability of reporting emissions for firms in the Communication sector (the baseline category) is approximately 17.9%. For firms in other sectors, the probability of disclosing emissions varies significantly. Specifically, we find that *brown* companies, on average, have a higher probability of disclosing emissions compared to *green* companies. In detail, firms in the Energy, Industrials, Materials and Utilities sectors have a reporting probability of approximately 76%, 83%, 84% and 79%, respectively. Conversely, firms in the Financials and Real Estate sectors have a lower probability of reporting emissions, with averages of 13% and 12%. These results suggest that companies subject to greater scrutiny from stakeholders or more stringent regulatory requirements tend to disclose more information, while those in less-polluting industries may face less pressure to disclose their environmental impact.

Additionally, a one standard deviation increase in firm size increases the probability of reporting to approximately 81.2%, reflecting that larger firms are significantly more likely to disclose their emissions. Conversely, a one standard deviation increase in Scope 1 emissions lowers the probability of reporting to approximately 43.8%, suggesting that firms with higher emissions are less likely to disclose their environmental performance. Notably, the coefficient associated with Scope 1 emissions displays higher variability than other covariates. This stems from the Monte Carlo procedure, in which random draws are rescaled by the standard deviation of each covariate. Despite this variability, the resulting probabilities remain stable: when accounting for the standard error of the emissions coefficient (± 0.0822), the probability of disclosure ranges from approximately 38.5% to 43.5%. This modest range reflects the non-linear transformation from log-odds to probabilities, where variations in the coefficient yield attenuated effects due to the curvature of the logistic function in this region.

Semi Parametric In contrast to the fully parametric approach, we also estimate the response propensity using a semi-parametric framework, specifically employing a Logistic Generalized Additive Model (LogisticGAM). This approach introduces multiple spline regressors, allowing for greater flexibility in capturing non-linear relationships between the covariates and the probability of disclosure, and without requiring any specification of the response model. However, this increased flexibility comes at the cost of interpretability, as the estimated coefficients vary across the range of the covariates.

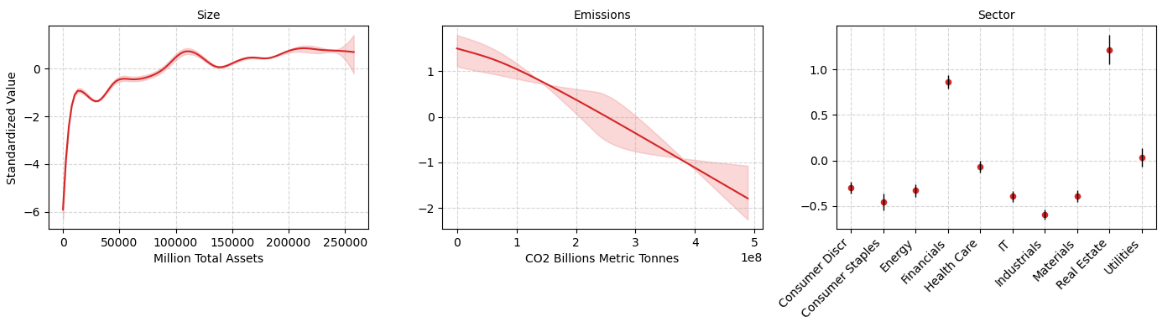

Figure 4: Distribution of standardized regression coefficients from the response mechanism model, based on 300 random draws under the semi-parametric specification. The left panel shows the estimated coefficients for firms' total assets; the center panel for Scope 1 emissions; and the right panel for business sector indicators. Red lines indicate the mean coefficient across the 300 iterations, while the lighter red bands, along with whiskers and outliers, illustrate the variability across draws.

Figure 4 presents the estimated standardized coefficient values for firm size, Scope 1 emissions, and the sector covariates, on the probability of emissions disclosure, along with their respective 95% confidence intervals. The left panel illustrates a non-linear relationship between firm size and disclosure. Specifically, the effect increases rapidly for small- to mid-sized firms and then stabilizes, indicating diminishing marginal effects at higher asset levels. Interestingly, firms in the lowest asset quantile exhibit a negative coefficient, suggesting that very small companies are less likely to disclose, potentially due to the fixed costs of reporting and limited regulatory/stakeholder pressure.

The middle panel highlights a clear, approximately linear negative relationship between Scope 1 emissions and disclosure likelihood. As emissions increase, the associated coefficient steadily declines, implying that firms with higher emissions are systematically less likely to disclose. This pattern is consistent with non-random disclosure, in which higher emitters are less likely to disclose, aligning with the MNAR framework adopted in the analysis.

The right panel displays the estimated sectoral effects relative to the baseline (Communication Services). Results align with fully parametric ones. Most sectors exhibit positive coefficients, indicating a higher probability of disclosure compared to the baseline. The most pronounced positive effects are observed for Energy, Industrials, and Consumer Staples, with the probability of reporting (with all other covariates constant) equal to 59%, 62%, and 61%, respectively. On the other hand, Financials and Real Estate report lower probabilities, equal to 29% and 24%, respectively.

Despite the advantages of the semi-parametric approach in capturing complex relationships, its interpretability remains limited due to the non-parametric nature of the splines. Nonetheless, the trends observed align with expectations and provide a nuanced understanding of the covariates' influence on disclosure probability, while validating the robustness of the fully parametric results.

6.3 Distributional effects of tilting

Building on the response mechanism discussed in the previous section, we now implement an adjustment of outcome model's predictions to correct for potential underreporting in companies' Scope 1 emissions. By leveraging forecasted response propensities, we adjusted the imputation process to better reflect the emissions of firms less likely to disclose. This approach allows us to account for systematic reporting biases and generate more credible estimates of environmental impact.

Under the assumption that firms with higher emissions are less likely to disclose them, we assign greater weights to observations with lower predicted disclosure probabilities when sampling from the response-weighted distribution, which tend to correspond to higher-emitting firms. The results discussed in this section are derived from random draws, as specified in Equation (7).

As described in Section 5, the tilting adjustment is derived from the response probability model estimations. The idea is that forecasted emissions from the outcome model are scaled based on the inverse likelihood of disclosure: the less likely a firm is

	B=300		B=500	
	Mean	St. Dev.	Mean	St. Dev.
MAR	1,229,200	95,345	1,224,434	103,729
MNAR (FP)	1,461,681	150,860	1,456,964	154,405
MNAR (SP)	1,412,524	143,221	1,408,151	153,358

Table 1: Summary statistics (mean and standard deviation) of Scope 1 emissions under different imputation strategies: non-tilted (based on the observable component of the data), and tilted (adjusted using fully or semi-parametric response models).

to report, the greater the upward adjustment. However, the final tilt does not depend only on disclosure probability but also on each company's size, business sector, and imputed emissions. Hence, a low reporting probability does not always result in a large adjustment.

Table 1 displays summary statistics of imputed Scope 1 emissions under different methodologies. When MAR is assumed, the average emissions equals about 1.2 million metric tonnes of CO_2. In contrast, when MNAR is assumed, the fully parametric and semi-parametric adjustments raise this average to 1.46 and 1.41 million metric tonnes, respectively. This represents a 19% increase for the fully parametric model and a 15% increase for the semi-parametric one, relative to the non-adjusted estimates. These results suggest that emissions may be substantially under-reported in the raw data.

6.4 Validation: Firms transitioning to disclosure

A fundamental challenge in settings with non-ignorable missingness is the absence of ground truth for validating imputed values. However, a subset of firms in our sample transition from non-disclosure to disclosure over time, providing a rare opportunity for external validation. Specifically, we identify 175 companies for which emissions were initially estimated, and then disclosed in subsequent years. For these firms, we compare emissions imputed in the year prior to disclosure with the subsequently reported values. This exercise does not identify the model, but offers a diagnostic assessment of whether MNAR-based corrections reduce systematic underestimation relative to standard MAR-based approaches.

While Trucost's estimates perform well on average, they systematically understate emissions in the upper tail. In contrast, the MNAR-based imputations substantially reduce underestimation among high emitters, which are precisely the firms that account for a disproportionate share of aggregate emissions.

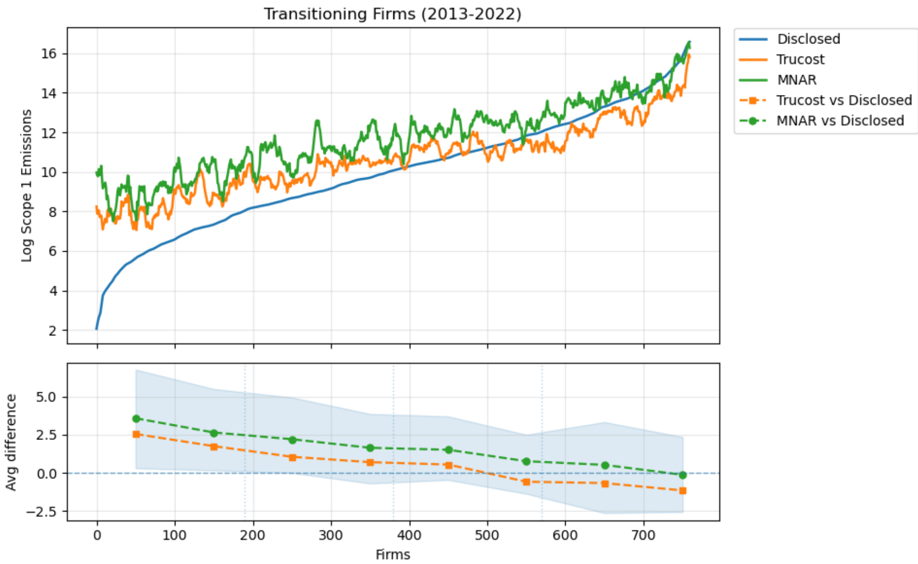

Figure 5: Comparison of estimated versus actual emissions for firms that transition to disclosure. Top: Predicted log Scope 1 emissions from Trucost (orange) and our MNAR-adjusted model (green) compared to subsequently disclosed values (blue). Bottom: Average difference between predictions and disclosures, in bins of 100 firms ranked by disclosed emissions. Shaded areas represent $\pm$ one standard deviation.

The lower panel of Figure 5 highlights this divergence. The average difference between Trucost's estimates and disclosed values becomes increasingly negative, indicating a growing tendency to underestimate emissions among the high emitters. By contrast, the differences for our MNAR-based model converge toward zero, suggesting improved alignment with ground truth in the upper tail. This pattern is consistent with our framework's design, which accounts for strategic non-disclosure incentives that are more prevalent among large emitters. Given that these firms contribute disproportionately to aggregate pollution, the reduction in high-end underestimation is critical for accurately estimating the total average emissions and associated carbon damages.

7 Implications

7.1 US Total Scope 1 Emissions

The imputation of missing data allows us to recover the full distribution of Scope 1 emissions across U.S. public companies. In 2022, we estimate that these companies emitted a total of 8,571.85 million metric tons of CO_2 (MtCO2). For context, the U.S. Environmental Protection Agency (EPA) reported total U.S. greenhouse gas emissions in 2022 of 6,343.2 MtCO2 equivalent, encompassing all sectors and greenhouse gases[1].

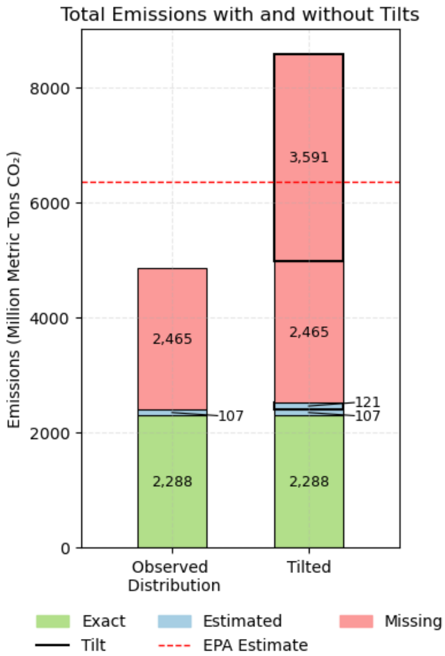

Figure 6: Total Scope 1 CO_2 emissions in 2022 under observed and tilt-adjusted scenarios. Bars are broken down into emissions that are *disclosed* by companies(green), *estimated* by Trucost (blue), and *missing* from Trucost but imputed with our estimated (red). Tilt adjustments are shown in bolder black lines. The dashed red line represents the EPA's estimate for U.S. corporate emissions.

Figure 6 presents the breakdown of total emissions by reporting type. We distinguish between firms that disclosed *exact* emissions (green); firms whose emissions are *estimated* by Trucost (blue); and firms for which emissions data are *missing*, i.e. companies in the Compustat dataset but not in Trucost Environmental (red). We observe that exact emissions data account for 2,288 MtCO2, while estimated and missing emissions contribute 228 and 6,056 MtCO2, respectively. Importantly, the figure illustrates the impact of tilt adjustments in our imputation approach. Among the 228 MtCO2 in estimated emissions, approximately 53% (121 MtCO2) result from tilting – an adjustment based on non-reporting probabilities. For the 6,056 MtCO2 classified as missing, 59% of emissions (3,591 MtCO2) arise from tilt-adjusted values. This pattern is consistent with non-random disclosure, whereby firms less likely to report tend to have higher imputed emissions.

Figure A1 reports average Scope 1 emissions for the three sub samples of our data. In

[1]The EPA's national inventory is compiled through the EPA's annual Inventory of U.S. Greenhouse Gas Emissions and Sinks, which tracks emissions by source, economic sector, and gas using activity data, emissions factors, and facility-level reports from the Greenhouse Gas Reporting Program. https://www.epa.gov/ghgemissions/sources-greenhouse-gas-emissions

2022, firms with exact disclosure averaged 1.85 million metric tons of Scope 1 CO_2, while those with Trucost-estimated emissions contributed an additional 52.6 thousands tons per firms. In contrast, firms with missing emissions, exhibited average emissions of 538 thousand tons under baseline imputation – using their observed distribution. Howrever, after tilt correction, these firms averaged 1.31 million metric tons of CO_2 (an increase equal to 783 thousands CO_2). This sharp increase reflects the non-random nature of non-disclosure. In fact, this pattern underscores that low-emitting firms are more likely to disclose, while high-emitting firms tend to remain silent, thereby distorting the average unless corrected. These results reconcile with Figure **??** discussed previously.

Figure A2 shows sectoral in the composition of Scope 1 emissions. The impact of tilt adjustments varies significantly across sectors. Exact emissions constitute the majority share for Industrials (67%), IT (52%) and Materials (47%), reflecting higher rates of disclosure in these industries. In contrast, Real Estate (1%), Financials (12%) and Utilities (16%) report the lowest share of exact data. We note that in sectors such as Consumer Discretionary, Materials, and Financials, tilt-adjusted emissions are concentrated within the estimated portion of data.

The Real Estate sector stands out with the most prominent tilt adjustment, followed by Energy and Health Care. This reflects sector-specific disclosure incentives and differences in firm size distributions. In Real Estate, only 1% of total emissions are disclosed, resulting in heavy tilt penalties due to systematic underreporting. Energy firms, as high-emissions companies, are also heavily adjusted to account for their pollution intensity. Lastly, Health Care firms – while often considered *green* – are typically large companies (representing 9% of the total market assets in 2022, behind Financials at 29%, and tied with Industrials and Utilities as 11%). Their size increases the tilt penalty despite their greener profile, as non-disclosure from large firms can obscure a substantial portion of emissions. These patterns further support the notion that reporting behavior is shaped by sector-specific incentives and regulatory expectations, rather than being random or uniform across industries.

Finally, Figure 7 compares total Scope 1 CO_2 emissions for U.S. public companies from 2012 to 2023 under different imputation strategies. Specifically, we contrast our approach—based on the observed emissions distribution combined with tilts for non-reporting firms—with conventional methods that impute missing data using the mean or median of sector-year distributions. In both conventional approaches, the imputations are weighted by company size (i.e., total assets). The procedure involves three steps: (1) dividing reported Scope 1 emissions by each company's total assets,

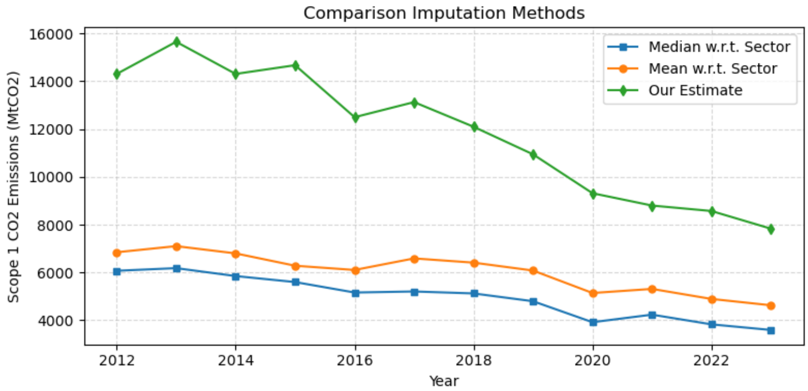

Figure 7: Comparison of total Scope 1 CO_2 emissions under different imputation strategies for U.S. public companies from 2012 to 2023. The figure contrasts our estimation method (green) with standard imputations based on the mean (orange) and median (blue) emission intensities by sector and year, weighted by firm size.

(2) computing the sector-year average (mean or median) of these emission intensities, and (3) estimating emissions for non-reporting firms by multiplying their assets by the corresponding sector-year intensity. The figure shows that both mean- and median-based imputations consistently underestimate emissions, although the gap narrows over time. Notably, in 2023, our estimate of Scope 1 emissions is approximately 70% and 120% higher than those derived from mean and median imputations, respectively. These results underscore the limitations of standard imputation techniques and the importance of accounting for firm-level disclosure behavior when estimating emissions. Importantly, these differences arise from how missingness is treated, not from differences in observed emissions among disclosing firms.

7.2 Corporate Carbon Damages

Greenstone et al. (2023) define corporate carbon damages as *a measure of the total societal costs resulting from corporate emissions*. Specifically, these damages are calculated as the product of a firm's direct CO_2 emissions and the social cost of carbon (SCC), which represents the monetary value of the damages associated with emitting an additional metric ton of CO_2. We evaluate three scenarios - (i) low, (ii) mid, and (iii) high cost - with corresponding SCC values of \$51, \$190, and \$250 per tCO_2, respectively. To contextualize the magnitude of these damages at the firm level, we normalize them by expressing carbon damages as a fraction of a firm's revenues. Thus, we presents results both in absolute and relative to firm revenues, for comparability.

Section 6 presents an average measure of Scope 1 CO_2 emissions for all companies

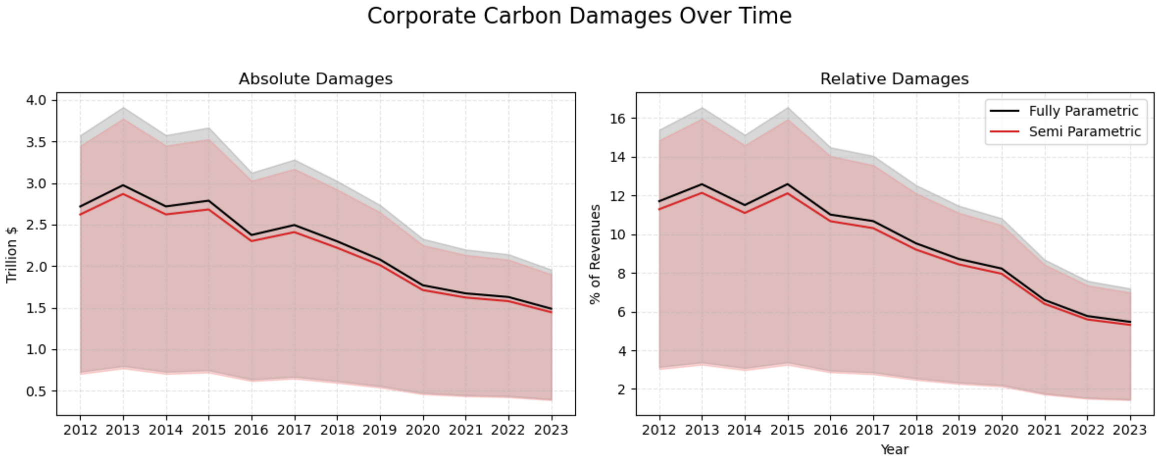

Figure 8: Evolution of U.S. carbon damages from 2012 to 2023. The left panel displays absolute damages in trillion USD, while the right panel shows damages relative to firm size (i.e., scaled by total assets). Estimates are based Greenstone et al. (2023) definition of corporate carbon damages scenarios: low (\$51/tCO_2), mid (\$190/tCO_2), and high (\$250/tCO_2). Solid red and black lines represent mid-cost estimates from the semi-parametric and fully parametric models, respectively. Shaded bands reflect the range between low- and high-cost scenarios.

in our sample, derived from 300 Monte Carlo random draws. This measure enables us to estimate the total carbon damages attributable to U.S. corporations. Figure 8 illustrates the evolution of carbon damages over time, comparing results from both fully and semi-parametric models. Consistent with our emissions forecasts, our model reveals a downward trend in carbon damages. Specifically, damages peaked in 2013 at USD 2.97 trillion under the mid-cost scenario, declining to USD 1.49 trillion by 2023. As a relative measure, these figures represented approximately 12.6% of total revenues in 2013 and 5.4% in 2023, respectively.

To contextualize our findings, we compare our relative carbon damage estimates with those forecasted by Greenstone et al. (2023) for the S&P 1500. Given that our sample includes over 8,000 companies per year—far exceeding the S&P 1500's scope—we focus on relative estimates rather than absolute values. For 2019, Greenstone et al. (2023) project average relative carbon damages of 0.5%, 2%, and 2.7% under the low-, mid-, and high-cost scenarios, respectively. In contrast, our fully parametric model yields higher estimates of 1.5%, 5.6%, and 7.2%, while the semi-parametric model produces 1.4%, 5.3%, and 7% for the same scenarios. These elevated figures are consistent with the greater heterogeneity in our sample, which encompasses a broader range of industries and firm sizes, leading to higher variability in emissions profiles. In fact, our broader sample captures a wider range of firm sizes and emissions profiles, extending beyond the S&P 1500.[2]. See Appendix A.5 for further details.

Cost Scenario	Low	Mid	High
Greenstone et Al.	0.5%	2%	2.7%
MNAR (FP)	1.5%	5.5%	7.2%
MNAR (SP)	1.4%	5.3%	7%

Table 2: Comparison of corporate carbon damages as a share of revenues under Greenstone et al. (2023) social cost of carbon (SCC) scenarios: low (\$51/tCO_2), mid (\$190/tCO_2), and high (\$250/tCO_2). The first row reproduces benchmark estimates from Greenstone et al. (2023). The second and third rows report our estimates using fully parametric and semi-parametric approaches based on imputed Scope 1 emissions. Results refer to average damages in 2019, expressed as a percentage of firm revenues.

Being able to impute single data points allows us to investigate the distribution of carbon damages across corporate sectors over the years. Figure A3 illustrate the allocation of carbon damages for fully and semi-parametric models, across all sectors, under the mid-scenario. As expected, brown sectors— Utilities, Energy, Materials, and Industrials—account for the highest carbon damages. These sectors inherently exhibit high carbon intensity due to the nature of their operations. Conversely, sectors such as Health Care, Communication, IT, and Consumer Staples report the lowest carbon damages, reflecting their relatively lower carbon footprints. Notably, brown sectors collectively contribute to more than 90% of total carbon damages, underscoring their dominant role in driving emissions. In detail, in 2023, Utilities, Materials, Energy and Industrials contributed to 60%, 12%, 17%, and 6% of total carbon damages, respectively. This concentration in a few high-emission sectors highlights the need for targeted emission reduction policies and sector-specific mitigation strategies to effectively reduce overall carbon damages.

8 Carbon Emissions and Expected Returns

The preceding sections show that Scope 1 emissions are selectively disclosed and that MNAR-based imputation substantially reshapes the inferred emissions distribution. In this section, we conduct a diagnostic exercise to assess whether these corrections materially alter empirical relationships in which emissions are commonly used as regressors. We focus on stock returns, not to re-evaluate the existence of a carbon premium, but to

[2]Note that this higher variability may stem from the inclusion of smaller firms in our sample. As discussed in Section 6, these companies are less likely to disclose emissions data, potentially introducing a right-tail bias in our forecasts.

verify that the imputation procedure preserves existing return–emissions relationships and does not generate spurious pricing effects.

Table 3: **Emissions Imputation and Stock Returns (Diagnostic).** Panel regressions of monthly stock returns on standardized Scope 1 CO_2 emissions. All specifications include firm characteristics, year fixed effects, and industry fixed effects. Columns compare regressions using observed emissions and MNAR-imputed emissions on the same firm–month samples. Standard errors are clustered at the firm level.

	Disclosed Sample	Estimated Sample	Combined Sample
Observed Emissions	0.024	0.062*	0.025
	(0.026)	(0.025)	(0.018)
MNAR-Imputed Emissions	0.021	-0.207	0.015
	(0.021)	(0.198)	(0.023)
Controls	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
Industry FE	Yes	Yes	Yes
Observations	56,391	95,866	152,257
R^2	0.299	0.250	0.260

Notes: Emissions are standardized within year. The table reports the coefficient on emissions only; all control variables are included but not shown. Significance levels: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

We estimate standard panel regressions of monthly stock returns on lagged firm characteristics and emissions, including year and industry fixed effects (in the spirit of Bolton and Kacperczyk (2021)). The analysis is conducted separately for firms with disclosed emissions, firms with vendor-estimated emissions, and the combined sample. We then replace observed emissions with their MNAR-imputed counterparts and re-estimate the same specifications. Table 3 reports the results.

Across all specifications, the estimated relationship between emissions and returns is economically small and statistically insignificant. Importantly, replacing observed emissions with MNAR-imputed values leaves the estimated coefficients essentially unchanged. This finding indicates that the imputation procedure preserves the underlying return–emissions relationship present in the data, rather than introducing artificial correlations.

In contrast, regressions that rely on vendor-estimated emissions produce materially different coefficients, consistent with measurement error in estimated data. The similarity between coefficients obtained using disclosed emissions and MNAR-imputed emissions provides external validation that the imputation framework corrects for missingness without distorting downstream empirical relationships.

9 Conclusion

Accurate corporate emissions data are essential for climate analysis, yet disclosure remains incomplete and systematically biased. In this paper, we tackle this problem by focusing on Scope 1 emissions among U.S. public companies from 2012 to 2023. We propose a novel framework for imputation under a missing not at random (MNAR) assumption, motivated by the strategic nature of non-disclosure and its potential correlation with unobserved emissions levels.

Our approach integrates outcome modeling, response mechanism estimation, and exponential tilting to correct for selection bias in the disclosure process. Empirical results indicate that MNAR-corrected estimates increase average aggregate Scope 1 emissions by 15–19% relative to unadjusted figures, with over half of missing emissions imputed above baseline levels. These findings suggest that non-reporting behavior is concentrated among firms with above-average emissions, consistent with strategic withholding.

Sectoral and size-based heterogeneity in disclosure patterns further supports the notion that missingness is endogenous. Larger firms and those in high-emission sectors are more likely to report, while disclosure rates remain low among smaller firms and those in sectors with less direct regulatory or reputational exposure. The uneven distribution of missingness underscores the importance of explicitly modeling the disclosure mechanism, rather than relying on ignorable missing data assumptions.

Finally, by applying estimates of the social cost of carbon, we show that the implications of under-reported emissions extend to assessments of corporate carbon damages. Our damage estimates are substantially higher than those implied by self-reported data, particularly for non-disclosing firms. This reinforces the broader relevance of our methodological contribution to both empirical research and applied environmental accounting.

Overall, our results demonstrate that addressing non-ignorable missingness is essential for generating credible emissions inventories. Future work could extend the methodology to include Scope 2 and Scope 3 emissions, explore firm-level dynamics over time, or evaluate the effects of regulatory interventions on disclosure incentives. As climate disclosure frameworks become more widely adopted and potentially mandated, the tools developed in this article provide a foundation for anticipating and correcting for persistent gaps in voluntary reporting.

References

Assael, J., Heurtebize, T., Carlier, L., & Soupé, F. (2023). Greenhouse gases emissions: Estimating corporate non-reported emissions using interpretable machine learning. *Sustainability*, *15*(4), 3391.

Aswani, J., Raghunandan, A., & Rajgopal, S. (2024). Are carbon emissions associated with stock returns? *Review of Finance*, *28*(1), 75–106.

Berg, F., Oliver Huidobro, J., & Rigobon, R. (2024). On the importance of assurance in carbon accounting. *Roberto, On the Importance of Assurance in Carbon Accounting (February 23, 2024)*.

Bolton, P., & Kacperczyk, M. (2021). Do investors care about carbon risk? *Journal of financial economics*, *142*(2), 517–549.

Cheema-Fox, A., LaPerla, B. R., Serafeim, G., Turkington, D., & Wang, H. (2021). Decarbonizing everything. *Financial Analysts Journal*, *77*(3), 93–108.

Chen, C. Y.-H., Lioui, A., & Scaillet, O. (2025). Green silence: Double machine learning carbon emissions under sample selection bias.

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.

Dhaliwal, D. S., Li, O. Z., Tsang, A., & Yang, Y. G. (2011). Voluntary nonfinancial disclosure and the cost of equity capital: The initiation of corporate social responsibility reporting. *The accounting review*, *86*(1), 59–100.

Downar, B., Ernstberger, J., Reichelstein, S., Schwenen, S., & Zaklan, A. (2021). The impact of carbon disclosure mandates on emissions and financial operating performance. *Review of Accounting Studies*, *26*(3), 1137–1175.

Duan, J., & Pelger, M. (2025). *Imputation-powered inference for missing covariates* (tech. rep.). National Bureau of Economic Research.

Dyck, A., Lins, K. V., Roth, L., & Wagner, H. F. (2019). Do institutional investors drive corporate social responsibility? international evidence. *Journal of financial economics*, *131*(3), 693–714.

Dye, R. A. (1985). Disclosure of nonproprietary information. *Journal of accounting research*, 123–145.

Fankhauser, S. (1994). The social costs of greenhouse gas emissions: An expected value approach. *The energy journal*, *15*(2), 157–184.

Gohil, K. (2020). Multiple imputation using influential exponential tilting in case of non-ignorable missing data.

Greenstone, M., Leuz, C., & Breuer, P. (2023). Mandatory disclosure would reveal corporate carbon damages. *Science*, *381*(6660), 837–840.

Grossman, S. J. (1981). The informational role of warranties and private disclosure about product quality. *The Journal of law and Economics*, *24*(3), 461–483.

Halkos, G., & Nomikos, S. (2021). Corporate social responsibility: Trends in global reporting initiative standards. *Economic Analysis and Policy*, *69*, 106–117.

Kott, P. S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, *32*(2), 133.

Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American statistical Association*, *83*(404), 1198–1202.

Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data*. John Wiley & Sons.

LSEG. (2024). *Lseg climate data package: Ghg emissions methodology and process* [Methodology document]. LSEG Data & Analytics. https://www.lseg.com/content/dam/data-analytics/en_us/documents/methodology/lseg-climate-data-package-ghg-emissions-methodology-process.pdf

Lyon, T. P., & Montgomery, A. W. (2015). The means and end of greenwash. *Organization & environment*, *28*(2), 223–249.

Milgrom, P. (1981). Good news and bad news: Representation theorems and applications. *Bell Journal of Economics*, *12*, 380–391.

Nguyen, Q., Diaz-Rainey, I., & Kuruppuarachchi, D. (2021). Predicting corporate carbon footprints for climate finance risk analyses: A machine learning approach. *Energy Economics*, *95*, 105129.

Olesiewicz, M., Kooroshy, J., & Greven, S. (2021). Navigating the corporate disclosure gap: Modelling of missing not at random carbon data. *arXiv preprint arXiv:2112.07784*.

Pastor, L., Stambaugh, R. F., & Taylor, L. A. (2025). Green tilts. *Journal of Financial Economics*, *174*, 104173.

Qin, J., Shao, J., & Zhang, B. (2008). Efficient and doubly robust imputation for covariate-dependent missing responses. *Journal of the American Statistical Association*, *103*(482), 797–810.

Rubin, D. B. (1987). *Statistical analysis with missing data*. Wiley.

Shi, Y., Tang, C. S., & Wu, J. (2025). Are firms voluntarily disclosing emissions greener? *Production and Operations Management*, *34*(8), 2363–2377.

S&P Global. (2025, November). *Trucost environmental data: Methodology* [Methodology guide]. S&P Global. https://portal.s1.spglobal.com/survey/documents/SPG_S1_Trucost_Environmental_Data_Methodology.pdf

Verrecchia, R. E. (1983). Discretionary disclosure. *Journal of Accounting and Economics*, *5*, 179–194.

Wang, S., Shao, J., & Kim, J. K. (2014). An instrumental variable approach for identification and estimation with nonignorable nonresponse. *Statistica Sinica*, 1097–1116.

Yang, S., & Kim, J. K. (2017). A semiparametric inference to regression analysis with missing covariates in survey data. *Statistica Sinica*, 261–285.

A Appendix

A.1 Sample Statistics

Table A1: **Sample Summary Statistics**. Data are extracted from Trucost Environmental and Compustat, for US public companies from 2012 to 2023. Data are yearly values. Total Market Value is in Billions USD. Total CO_2 Emissions are Scope 1 Emissions, measured in Metric Tons, both disclosed by companies and estimated by Trucost. Average CO_2 intensity is computed as total emissions per total assets of each company, in Tons of CO_2 over Millions USD.

Year	N	Total Mkt Value	Scope 1 Emissions	Intensity[3]
2012	9419	18,831	2.47	279.3
2013	9376	23,985	2.52	271.9
2014	9123	25,665	2.60	263.1
2015	8837	24,714	2.46	263.0
2016	8563	26,585	2.58	168.7
2017	8299	31,214	2.69	165.7
2018	8154	30,148	2.76	156.5
2019	8222	34,686	2.7	163.6
2020	8318	41,210	2.56	160.8
2021	8283	52,289	2.73	148.5
2022	8112	41,526	2.64	107.9
2023	7600	50,076	2.47	102.7

[3] CO_2 Emission divided by total assets of a company (CO_2 Metric Tonnes per $Million).

Table A2: **Sources of Scope 1 emissions data, 2023 snapshot**.

Disclosure source	**Count**	**Share (%)**	**Treatment**
Vendor-estimated emissions			
Estimated data	2,044	26.9	Missing
Estimate based on partial data disclosure (CSR)	2	<0.05	Missing
Estimate based on partial data disclosure (Annual report / 10-K)	1	<0.05	Missing
Firm-reported emissions (CSR / Environmental reports)			
Value derived from data provided in Environmental/CSR	784	10.3	Observed
Exact value from Environmental/CSR	118	1.6	Observed
Value summed up from data provided in Environmental/CSR	32	0.4	Observed
Value split from data provided in Environmental/CSR	9	0.1	Observed
Value derived from fuel use provided in Environmental/CSR	12	0.1	Observed
Derived from previous year	189	2.5	Observed
Firm-reported emissions (CDP)			
Exact value from CDP	54	0.7	Observed
Value derived from data provided in CDP	31	0.4	Observed
Firm-reported emissions (Annual reports / 10-Ks)			
Value derived from data provided in Annual Report / Financial Accounts	24	0.3	Observed
Exact value from Annual Report / 10-K / Financial Accounts	3	<0.05	Observed
Value derived from fuel use provided in Annual Report / Financial Accounts	6	<0.05	Observed
Value summed up from data provided in Annual Report / Financial Accounts	1	<0.05	Observed
Value split from data provided in Annual Report / Financial Accounts	1	<0.05	Observed
Total	**3,311**	100.0	

Table A3: **Summary statistics by disclosure bins**. Data are extracted from Trucost Environmental and Compustat, for US public companies from 2012 to 2023. Data are yearly values. Total Market Value is in Billions USD. Total CO_2 Emissions are Scope 1 Emissions, measured in Metric Tons, both disclosed by companies and estimated by Trucost. Average CO_2 intensity is computed as total emissions per total assets of each company, in Tons of CO_2 over Millions USD.

Year	N	Market cap	Scope 1 emissions	CO_2 intensity
Panel A: Observed emissions				
2012	371	10.02	2.30	485.50
2013	394	12.94	2.34	469.82
2014	419	14.35	2.41	444.09
2015	435	14.16	2.21	426.21
2016	483	16.45	2.30	475.66
2017	525	19.16	2.21	451.39
2018	593	18.88	2.35	403.19
2019	672	23.38	2.34	394.93
2020	805	28.14	2.16	407.00
2021	1,005	38.95	2.29	327.94
2022	1,244	33.54	2.40	214.64
2023	1,221	39.10	2.13	208.72
Panel B: Missing emissions				
2012	9,048	8.81	0.17	80.67
2013	8,982	11.04	0.17	102.34
2014	8,704	11.31	0.18	95.33
2015	8,402	10.55	0.25	108.30
2016	8,080	10.13	0.29	81.39
2017	7,774	12.05	0.48	81.35
2018	7,561	11.27	0.42	75.82
2019	7,550	11.31	0.36	77.52
2020	7,513	13.07	0.41	65.07
2021	7,278	13.34	0.44	58.22
2022	6,868	7.99	0.24	44.23
2023	6,379	10.97	0.34	40.79

A.2 Outcome Model Results

Table A4: **Outcome model results.** Linear regression of log Scope 1 emissions on firm characteristics, macroeconomic conditions, sector indicators, and year fixed effects. Standard errors in parentheses. Robust standard errors in parentheses. Significance levels: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

	Log Scope 1 Emissions
Macroeconomic and firm characteristics	
GDPC1	0.0497**
	(0.020)
(log) Assets	0.9969***
	(0.015)
Constant	0.0866***
	(0.017)
Sector fixed effects (baseline: Comm. Services)	
Consumer Discretionary	1.8044***
Consumer Staples	3.0524***
Energy	4.9035***
Financials	-2.7083***
Health Care	0.6620***
IT	0.4873***
Industrials	2.6329***
Materials	4.6575***
Real Estate	0.7314**
Utilities	5.4243***
Year fixed effects (baseline: 2012)	
2013	-0.0163
2014	-0.0779
2015	-0.0825
2016	-0.1669
2017	-0.2027
2018	-0.2681**
2019	-0.3776***
2020	-0.5030***
2021	-0.6683***
2022	-0.7097***
2023	-0.8249***
Observations	8,064
R^2	0.672

A.3 Response Mechanism Results

Table A5: **Response mechanism: fully parametric model.** Mean coefficient estimates and standard errors from Monte Carlo simulations. Results are reported for $B = 300$ and $B = 500$ draws.

	$B = 300$		$B = 500$	
	Mean	SE	Mean	SE
Firm characteristics				
Constant	-1.522	0.010	-1.522	0.010
Size	1.526	0.020	1.526	0.020
Emissions	-0.301	0.082	-0.540	0.142
Sector effects (baseline: Communication Services)				
Consumer Discretionary	1.257	0.005	1.257	0.005
Consumer Staples	1.501	0.005	1.501	0.005
Energy	1.108	0.013	1.109	0.012
Financials	-0.973	0.005	-0.973	0.005
Health Care	0.200	0.005	0.200	0.005
IT	1.135	0.006	1.136	0.006
Industrials	1.613	0.007	1.613	0.007
Materials	1.053	0.012	1.053	0.012
Real Estate	-1.053	0.003	-1.053	0.003
Utilities	1.324	0.047	1.325	0.046

Notes: Coefficients are standardized. Results are based on repeated simulations of the response mechanism estimated on imputed emissions.

A.4 Estimated Scope 1 Emissions

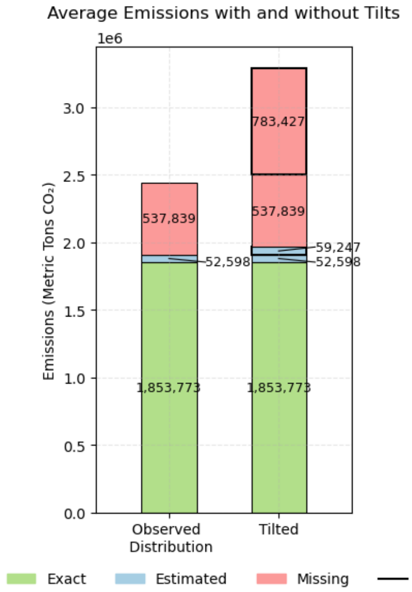

Figure A1: **2022 Average Scope 1 CO_2 emissions** under observed and tilt-adjusted scenarios. Bars are broken down into emissions that are *exactly* disclosed (green), *estimated* by Trucost (blue), and *missing* from Trucost but imputed from Compustat data (red). Tilt adjustments are shown in bolder black lines.

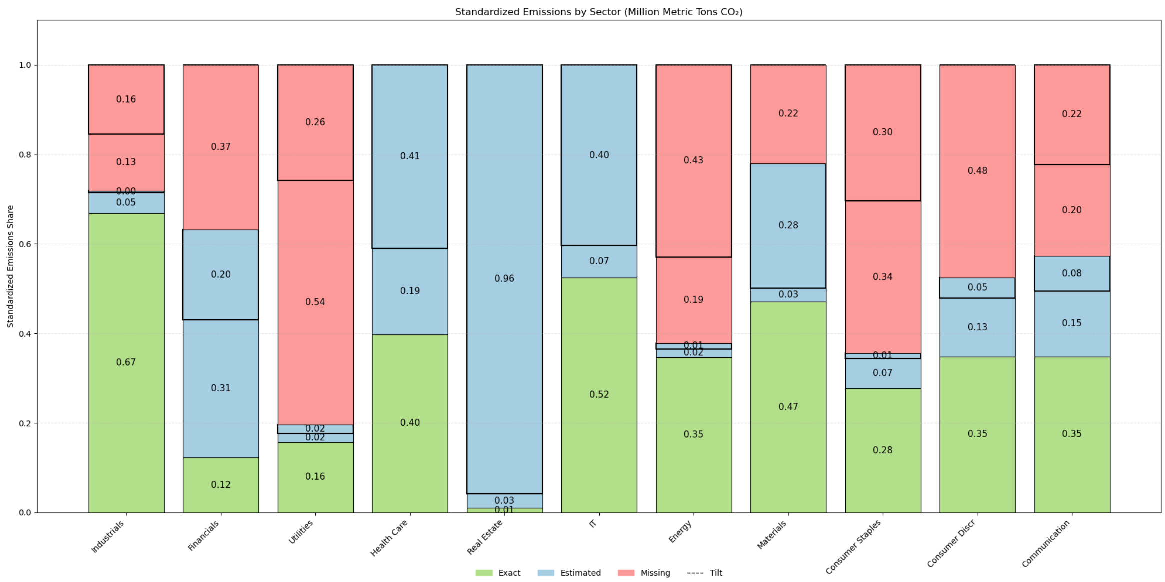

Figure A2: **2022 Percentage breakdown of Total Scope 1 CO_2 emissions** under observed and tilt-adjusted scenarios, for each business sectors. Bars are broken down into emissions that are *exactly* disclosed (green), *estimated* by Trucost (blue), and *missing* from Trucost but imputed from Compustat data (red). Tilt adjustments are shown in bolder black lines.

A.5 Corporate Carbon Damages Evolution

Table A6: **Absolute Carbon Damages**. Evolution of U.S. carbon damages from 2012 to 2023, expressed in trillion USD. Estimates are based on Scope 1 emissions and social cost of carbon (SCC) values drawn from Greenstone et al. (2023) scenarios: low (\$51/tCO_2), mid (\$190/tCO_2), and high (\$250/tCO_2).

	Fully Parametric			Semi-Parametric		
Year	Low	Mid	High	Low	Mid	High
2012	0.73	2.72	3.58	0.70	2.62	3.45
2013	0.80	2.97	3.91	0.77	2.87	3.77
2014	0.73	2.72	3.58	0.70	2.62	3.45
2015	0.75	2.79	3.67	0.72	2.68	3.53
2016	0.64	2.37	3.12	0.62	2.30	3.03
2017	0.67	2.49	3.28	0.65	2.41	3.17
2018	0.62	2.30	3.02	0.60	2.22	2.92
2019	0.56	2.08	2.74	0.54	2.01	2.65
2020	0.48	1.77	2.33	0.46	1.71	2.25
2021	0.45	1.67	2.20	0.44	1.62	2.13
2022	0.44	1.63	2.14	0.42	1.58	2.08
2023	0.40	1.49	1.96	0.39	1.45	1.90

Table A7: **Relative Carbon Damages**. Evolution of U.S. relative carbon damages from 2012 to 2023, relative to firm size (i.e., scaled by total assets).

	Fully Parametric			Semi-Parametric		
Year	Low	Mid	High	Low	Mid	High
2012	3.14	11.70	15.40	3.03	11.29	14.85
2013	3.38	12.58	16.56	3.26	12.14	15.97
2014	3.09	11.50	15.14	2.98	11.10	14.60
2015	3.38	12.59	16.57	3.25	12.11	15.93
2016	2.96	11.01	14.49	2.86	10.67	14.04
2017	2.87	10.68	14.05	2.77	10.31	13.57
2018	2.56	9.53	12.53	2.47	9.20	12.11
2019	2.34	8.72	11.47	2.26	8.44	11.10
2020	2.21	8.22	10.82	2.14	7.95	10.47
2021	1.77	6.60	8.69	1.72	6.40	8.43
2022	1.55	5.77	7.59	1.50	5.59	7.36
2023	1.47	5.47	7.20	1.43	5.31	6.99

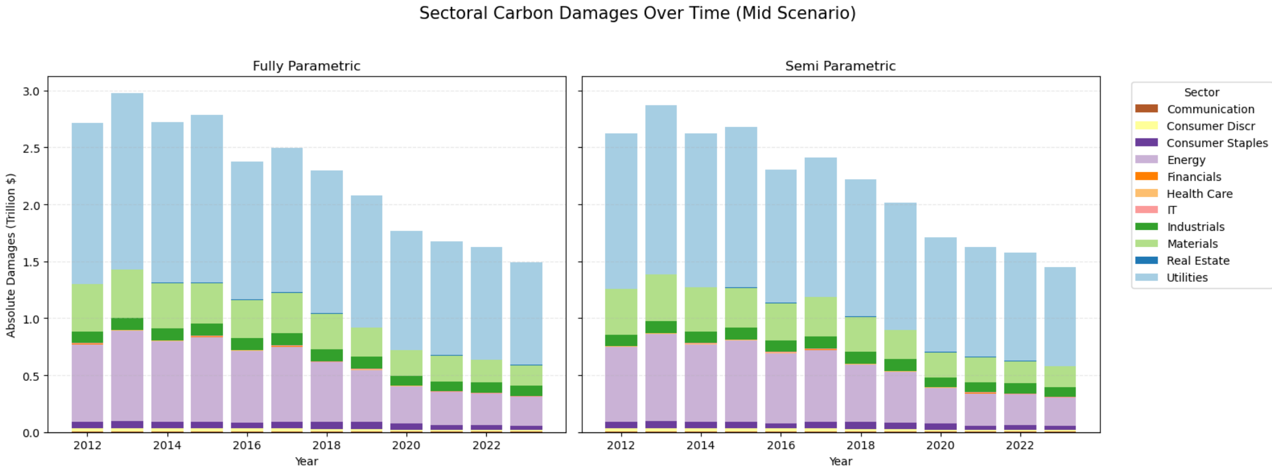

Figure A3: **Sectoral Carbon Damages**. Absolute corporate carbon damages by sector from 2012 to 2023. Estimates are based on imputed Scope 1 emissions under the mid-cost scenario (social cost of carbon = \$190/tCO_2), following Greenstone et al. (2023). Bars show sector-level averages using the fully parametric (left graphd) and semi-parametric methods (right graph).