Reading the Modeled–Disclosed Gap: Why Scope 3 Accuracy Depends on Disclosure Completeness

J. Christopher Hughen

June 27, 2026

Abstract

Environmentally-extended input–output (EEIO) models supply most of the Scope 3 emission estimates that investors and regulators now treat as data. Whether they match what firms actually disclose is untested at scale. S&P Global Trucost reports both a modeled and a disclosed value for each firm-year. Holding the provider fixed isolates model error from the cross-vendor noise that confounds prior comparisons. Across 266,929 firm-years (2005–2024), modeled figures track disclosed Scope 1 and location-based Scope 2 in rank (Spearman 0.86 and 0.93). For those scopes, though, the modeled series adopts the disclosed value where firms report, so the agreement holds largely by construction. Upstream Scope 3 is the cleaner test, since there the modeled figure is independent of disclosure. It diverges sharply: a rank correlation of 0.69, a log mean absolute error of 1.66, a median discrepancy near fivefold. Most of that gap is mechanical, governed by how completely a firm discloses. A gradient-boosted correction cuts out-of-sample error against disclosures by 11 to 15%, while a linear one does not. The residual error is structural, not a taxonomy artifact: half of intensity variation sits within sector, and neither SICS nor SASB materiality organizes accuracy any better. For users, the lesson is practical. Where firms disclose, that disclosure should anchor the estimate; absent it, EEIO is best treated as a rank and corrected when a level is required.

Keywords: Scope 3 emissions; environmentally-extended input–output analysis; hybrid life-cycle assessment; carbon data quality; greenhouse-gas disclosure; machine learning.

J. Christopher Hughen, Reiman School of Finance, University of Denver.

1 Introduction

For most firms, value-chain (Scope 3) emissions exceed direct (Scope 1) and energy-related (Scope 2) emissions, often by an order of magnitude (World Resources Institute and World Business Council for Sustainable Development, 2011; Hertwich and Wood, 2018). They are also the emissions on which climate-disclosure rules, financed-emissions accounting, and science-based target validation increasingly turn (Partnership for Carbon Accounting Financials, 2020; Science Based Targets initiative, 2023). Most firms do not report Scope 3, and those that do rarely report it in full. Commercial data providers fill the gap by supplying estimated Scope 3 figures for tens of thousands of firms. The dominant estimation technology is environmentally-extended input–output (EEIO) analysis, which routes a firm's revenue through sector emission intensities to approximate its value-chain footprint (Kitzes, 2013; Minx et al., 2009).

These estimates are used as if they were measurements. They feed portfolio screens and climate stress tests, and they sit at the core of the financed-emissions inventories that banks and asset managers now report. How close are the modeled numbers to what firms, when they do report, actually disclose? The evidence is surprisingly thin. A growing literature shows that carbon and ESG data diverge across providers (Busch et al., 2022; Berg et al., 2022) and that revenue-based input–output estimates are coarse at the firm level (Steen-Olsen et al., 2014; Su et al., 2010; Matthews et al., 2008). That work either compares providers to one another or sets the vendor's estimate aside as a contaminant in favor of disclosures (Aswani et al., 2024). Where a modeled figure has been set against a reported one, the two have come from different vendors, so model error is confounded with cross-vendor methodology (Nguyen et al., 2023). A direct, at-scale comparison of a single provider's modeled value against that same firm's disclosure has not been assembled.

We use a feature of one widely used provider, S&P Global Trucost, to build that comparison. For each firm-year Trucost records both an EEIO-modeled value and, where the firm has reported one, a company-disclosed ("as reported") value, scope by scope. Comparing the two within a single provider isolates the model's error from the cross-vendor methodology noise that confounds provider-to-provider studies. Assembling a global panel of 266,929 firm-years over 2005–2024, we address six questions:

RQ1. How closely do the modeled figures track disclosures, in level and in rank?
RQ2. Where is the gap largest across scopes, firm size, sector, and time?
RQ3. Which firm and sector characteristics are associated with a larger gap?
RQ4. Can a model that learns from financial and sector structure improve on the EEIO estimate?
RQ5. Does a sustainability-native industry taxonomy (SICS) organize that accuracy better than the market-native GICS on which the estimate is built?
RQ6. Does SASB's judgment that GHG emissions are financially material to an industry predict where the estimate is accurate, or how completely firms disclose?

Four findings follow. The model tracks disclosed direct emissions well in rank, yet it diverges sharply for upstream Scope 3, where the modeled figure is genuinely independent of the disclosure. There the within sector–year rank correlation is 0.69 and the mean absolute error is 1.66 in natural logarithms, a median multiplicative discrepancy of about $e^{1.66} \approx 5.3$. That magnitude is the firm-level, modeled-versus-disclosed analogue of the roughly fourfold gap Klaaßen and Stoll (2021) report when they harmonize the Scope 3 of a set of large technology firms.

Disclosure completeness drives the second result. Although the incompleteness of Scope 3 reporting is itself well documented (Klaaßen and Stoll, 2021; Nguyen et al., 2023), we show that it carries a specific consequence for any modeled-versus-disclosed comparison. Because a disclosed total can only sum the categories a firm reports, the apparent gap is mechanically governed by how

completely that firm discloses. Treating disclosure breadth, the count of reported categories, as the conditioning variable, we find it dominates the cross-sectional variation in the gap. Once we condition on breadth, firm fundamentals add little, with supply-chain reach the main exception.

The correction is the third finding. A gradient-boosted model with a log link reduces out-of-sample error against disclosures by 11 to 15% relative to the pure input–output estimate. Earlier machine-learning work predicts emissions to gap-fill non-disclosers (Nguyen et al., 2021; Serafeim and Velez Caicedo, 2022); we instead benchmark the correction against the EEIO figure a user would otherwise adopt. An interpretable linear version does not improve on it, so the useful correction is nonlinear. Those gains concentrate in sectors whose supply chains depart from the revenue-intensity assumption.

Aggregation rewards the right transformation. Modeling each value-chain category on its own reveals which parts are learnable from financial signals, and reassembling the predicted levels beats a direct model of the total. This matches what Nguyen et al. (2023) find in a cross-provider setting. It holds only because we model levels and sum them, which sidesteps the retransformation bias a log-space reassembly would incur.

A fifth question concerns the lens itself. Every statistic above is computed within sectors, yet the modeled figure is itself a sector construct. Trucost routes revenue through industry emission intensities defined on GICS, so a firm's EEIO estimate is close to a GICS-industry average scaled by sales. GICS, like the schemes that dominate capital-market research (Bhojraj et al., 2003), groups firms by product market and revenue line rather than by the resource intensity that drives emissions. The Sustainable Industry Classification System (SICS), developed by SASB and now maintained within the IFRS Foundation, instead groups firms by shared sustainability risk and resource intensity (Sustainability Accounting Standards Board, 2023). That is the very dimension EEIO holds constant within a sector. If disclosed emissions cluster more tightly within SICS cells than within GICS cells, the input–output homogeneity assumption is mis-specified against a sustainability-native partition. The apparent firm-level error would then be, in part, a wrong-bucket aggregation artifact (Steen-Olsen et al., 2014; Su et al., 2010; Zhang et al., 2019). We take up this question in Section 5, re-evaluating the modeled-versus-disclosed gap under SICS and GICS at matched granularity. The answer is a clean null. SICS and GICS leave almost identical within-sector dispersion in emission intensity, and SICS adds nothing even on the firms it deliberately regroups. On the rank agreement that portfolio screening relies on it is marginally worse. The firm-level error is therefore intrinsic to revenue-based estimation, not an artifact of the market-based taxonomy EEIO happens to use. That forecloses an obvious objection, that a sustainability-native classification would dissolve the gap, and it is, to our knowledge, the first test of SICS against the accuracy of estimated emissions. Comparisons of classification schemes have concerned stock returns and valuation (Bhojraj et al., 2003), and prior uses of SICS have addressed materiality and ESG ratings (Khan et al., 2016) rather than modeled-versus-disclosed error (Busch et al., 2022).

A sixth question asks whether financial materiality, not just the taxonomy, marks where the estimate can be trusted. SASB's materiality map flags the industries for which GHG emissions are financially material, the industries on which financed-emissions accounting and carbon-risk pricing concentrate. The estimate is in fact more accurate there: in GHG-material industries the modeled figure is nearly unbiased, while in the rest it overstates disclosed upstream emissions by more than twofold. That reassurance is real, but the mechanism is not materiality. Once we condition on the firm's direct carbon intensity, the materiality advantage disappears. A placebo across six other SASB issues confirms why. The apparent accuracy gain tracks how carbon-intensive an issue's industries are, not whether the issue concerns emissions, and it vanishes for every issue once intensity is held fixed. Materiality also fails to predict disclosure, neither whether a firm reports

upstream Scope 3 nor how completely, once size and region are fixed. Like the taxonomy null, this locates the firm-level structure in continuous fundamentals rather than in a sustainability-native label.

None of this is a case against EEIO estimation, which remains a reasonable default where firms are silent and is close to right on average. The results instead locate where it is least reliable at the firm level and where disclosure or a hybrid correction adds the most value. Sections 2–4 situate the study and lay out the data and methods; Section 5 presents the results; and Sections 6–7 discuss implications and limitations.

2 Related work

This study draws on three strands of research that have largely developed apart. They are the input–output accounting of emissions, the empirical study of commercial carbon-data quality, and the use of machine learning to estimate corporate emissions.

Input–output and hybrid life-cycle accounting. Environmentally-extended input–output analysis descends from Leontief's environmental extension of the input–output model (Leontief, 1970) and has become a workhorse for carbon footprinting at national, sectoral, and product levels (Minx et al., 2009; Kitzes, 2013). Its appeal is coverage: given a firm's revenue and a sector code, it returns an estimate of the entire upstream footprint without firm-specific activity data. That same strength is its weakness at the firm level, because it assigns every firm in a sector the same emission intensity. Aggregation error in the resulting input–output multipliers is large, well quantified, and sensitive to how sectors are grouped (Steen-Olsen et al., 2014; Su et al., 2010; Zhang et al., 2019; Piñero et al., 2015). Life-cycle researchers have long developed hybrid methods that graft firm- or process-specific data onto the input–output backbone to reduce that error (Suh and Huppes, 2005; Crawford et al., 2018). Closest to our question, a recent multiregional-accounting study benchmarks single-region EEIO upstream estimates against a multiregional alternative for CDP-reporting firms. The study finds the single-region figures understate aggregate upstream emissions by roughly a tenth (Davis et al., 2026). This work and ours are complementary but distinct. It treats a more detailed model as the reference and *corrects* the estimate at the aggregate level. We instead treat the firm's own disclosure as the reference and *audit* the estimate firm by firm.

Carbon- and ESG-data quality and divergence. A second literature documents how far apart commercial sustainability data can be. ESG ratings diverge substantially across providers, with much of the disagreement traceable to measurement rather than weighting (Berg et al., 2022). For carbon data specifically, providers disagree on emissions, restate aggressively, and backfill estimates in ways that can distort historical analysis (Busch et al., 2022; Kalesnik et al., 2022). This work establishes that the numbers users rely on are noisy. Yet it largely compares providers to one another or studies one series in isolation, and less often asks how an estimated figure compares to the disclosure for the same firm. In the work closest to that question, Aswani et al. (2024) show that vendor-estimated emissions behave statistically differently from firm-disclosed emissions and caution against treating the two as interchangeable.[1] Their concern is asset pricing, however, rather

[1]The asset-pricing reading of this distinction is debated. Aswani et al. (2024) relate the cross-section of returns to vendor-estimated rather than firm-disclosed emissions, challenging the carbon premium of Bolton and Kacperczyk (2021). In a published comment, Bolton and Kacperczyk (2024) dispute that reading and report a premium for disclosed emissions as well. That dispute concerns whether a carbon premium is priced, on which we take no position. We rely only on the point both sides accept: a revenue-based vendor estimate is mechanically tied to financial fundamentals. That is why we benchmark against disclosed rather than vendor-estimated values.

than the accuracy of EEIO Scope 3. We add the direct firm-year comparison for upstream Scope 3, within a single provider so that model error is not confounded with cross-vendor methodology.

Scope 3 disclosure. Disclosed Scope 3 is itself a noisy reference. Reporting is sparse, boundaries vary, and category coverage is uneven (Klaaßen and Stoll, 2021; Li et al., 2024). The magnitudes are large: harmonizing the Scope 3 of a set of major technology firms roughly doubles to quadruples their reported totals (Klaaßen and Stoll, 2021). That aggregate factor is the counterpart of the firm-level discrepancy we document below. Prior work has gestured at the consequence for comparison: Nguyen et al. (2023) note that incompleteness inflates apparent differences between reporters, and Han et al. (2021) frame disclosure-driven, "patterned" missingness as a dataset-shift problem. Our contribution is to make disclosure breadth, the category count, the conditioning variable for the accuracy metric. We use it to show that breadth mechanically governs the modeled-versus-disclosed gap, and that it dominates the firm fundamentals.

Machine learning for emissions estimation. A growing body of work applies statistical learning to predict corporate emissions from financial and sector data (Goldhammer et al., 2017; Han et al., 2021; Serafeim and Velez Caicedo, 2022; Nguyen et al., 2023). Serafeim and Velez Caicedo (2022) train on disclosed Scope 3 with financial variables, Scope-1/2 levels, and industry codes, and report sizable gains over linear baselines. Their benchmark, though, is ordinary least squares rather than an EEIO incumbent, drawn from a single, U.S.-centric vendor. Our emphasis differs in two ways. We benchmark the learned estimate against the *incumbent EEIO figure* on identical out-of-sample firm-years, using firm-grouped and forward-in-time validation to rule out leakage. The quantity of interest is therefore the improvement over the input–output number a user would otherwise adopt. Because a financials-based model could in principle re-derive a revenue-based vendor estimate (Aswani et al., 2024), we measure improvement against the *disclosed* value rather than the vendor's modeled value. That choice rules out a mechanically circular "win."

On the per-category question we reach the same conclusion as Nguyen et al. (2023), who find that modeling categories and aggregating improves on a direct model. In our setting the reassembled total likewise beats a direct model of the total, by a few percentage points of out-of-sample RMSE. Reaching that result depends on how the aggregate is formed. Each category is modeled as a level through a log link, and we sum the predicted levels, so the total carries no retransformation bias. Had we instead modeled categories in logs and exponentiated before summing, independent per-category errors would compound and a Jensen bias would enter. That alone is enough to flip the comparison and make reassembly look worse. Nguyen's setting still differs from ours, with a reference constructed across providers and an aggregate from a piecewise-linear forest. We read the agreement as evidence that per-category reassembly is viable once the summation is done on levels. The category view then serves twice over, locating the learnable signal and building a competitive aggregate.

How this study differs from the closest work. The nearest single study is Nguyen et al. (2023). It overlaps ours on three points: a large modeled-versus-reported Scope 3 divergence, a measure of Scope 3 disclosure incompleteness, and category-before-aggregate modeling. Three differences are nonetheless sharp. First, its comparison is across providers, modeled figures from one vendor against reported figures from others, so its gap conflates model error with cross-vendor methodology. Ours holds the provider fixed and isolates the model. Second, it records incompleteness as a fact, whereas we make the category count the conditioning variable and show it dominates the cross-sectional gap. That changes how any such comparison must be read. Third,

its machine-learning gains are measured against its own baseline. We instead benchmark against the incumbent EEIO figure a user would otherwise adopt, so our 11-to-15% is the improvement EEIO leaves on the table. The within-provider, modeled-versus-disclosed design has a precedent in Aswani et al. (2024), but in the service of asset pricing, not the accuracy of upstream Scope 3. A machine-learning lineage that predicts corporate footprints (Nguyen et al., 2021) gap-fills emissions for non-disclosers, a different target from auditing the estimate against the firm's own disclosure.

3 Data

3.1 A provider that reports both a modeled and a disclosed value

Our benchmark rests on a feature of the S&P Global Trucost database that is unusual among commercial carbon datasets. For each firm-year it records, side by side, a *modeled* emissions figure and, where the company has reported one, a *disclosed* ("as reported") figure. A source field documents the provenance of each value. The modeled figures are built on an environmentally-extended input–output (EEIO) framework that maps a firm's revenue, through sector emission intensities, into estimated emissions across the value chain. Disclosed figures are drawn from company channels (CDP responses, sustainability and annual reports) and reproduced at the level the firm itself published.

We access Trucost through Wharton Research Data Services. For every firm-year the data contain Scope 1 and Scope 2 (on both location- and market-based conventions), a Scope 3 upstream total, and Scope 3 reported at the category level. The category split follows the Greenhouse Gas Protocol's fifteen-category taxonomy (upstream categories 1–8, downstream categories 9–15). To confirm how each headline figure is produced, we inspected the source field directly. Provenance for the Scope-3 upstream total is almost entirely (about 99% of firm-years) "estimate derived by applying revenue-based intensity factors to revenue data," which is a pure EEIO estimate. By contrast, the Scope-1 and Scope-2 "modeled" fields adopt the disclosed value whenever a firm discloses, so for those scopes the modeled and disclosed series largely coincide. This distinction matters for interpretation, and we return to it below. The cleanest test of the input–output model is upstream Scope 3, where the modeled figure is genuinely independent of the firm's disclosure.

3.2 Constructing the comparison

The provider does not publish a single "upstream total as reported." We build the disclosed Scope-3 upstream total by summing the eight reported upstream categories (GHG Protocol categories 1–8, plus the small "other upstream" bucket). For each firm-year, we record how many of those categories carry a reported value. This completeness count is central to the analysis. A firm that reports only purchased goods will appear to "disclose" far less than the model estimates for its entire upstream chain. Any naive comparison of a coverage-complete modeled total against a partial disclosed sum then conflates model error with disclosure breadth. Throughout, we carry the count and condition every comparison on it.

We treat an emissions value of exactly zero as missing rather than as a measured zero, since a reported zero in these fields denotes non-coverage rather than genuinely zero emissions. Errors are computed in natural logarithms, so that they are multiplicative and comparable across firms of very different size.

3.3 Sample, identifiers, and covariates

The panel spans 2005–2024 and contains 266,929 firm-years for 44,860 listed firms with a `gvkey` link. Disclosed Scope 3 begins in earnest in 2020, so the modeled-versus-disclosed Scope-3 upstream comparison rests on 16,444 firm-years. Table 1 summarizes the composition. Emissions are in tonnes of carbon dioxide equivalent and revenue in U.S. dollars. Within Trucost the linking key is the institution identifier; we map it to `gvkey` to attach financial data.

We require firm financials both as covariates (firm size, supply-chain breadth, research and capital intensity, leverage, profitability) and as denominators. Because the disclosing universe is roughly four-fifths non-U.S., Compustat North America alone covers fewer than a third of the comparison firm-years. To extend coverage, we union Compustat North America and Global and, for the firm-years that remain unmatched, fill from Worldscope by bridging `gvkey` to ISIN to the Worldscope identifier. The fill is applied at the level of the firm-year, so a given firm-year takes all of its financials from one source and ratios never mix conventions. Where the two sources overlap they agree closely: the correlation of cost-of-goods-to-revenue is 0.75, with a median absolute difference of 0.008. Financial coverage on the comparison sample rises from about 86% to 99%. A size control built from the data must not be endogenous: a Trucost-implied revenue would share inputs with the EEIO model it is meant to evaluate. Real Compustat and Worldscope revenue and assets are therefore converted to dollars and used as the size measures in the financial specifications.

Finally, we construct a data-availability lag as the interval between a firm's fiscal year-end and the earliest date Trucost extracted the figure (median about 1.2 years). It is a proxy for disclosure timing rather than a filing date: it reflects both when a company disclosed and when the provider processed the record. We use it as a heterogeneity dimension and as a covariate, with that caveat stated.

3.4 Two industry classifications: GICS and SICS

The analysis conditions almost everything on sector, so the classification is not incidental. Our workhorse is the GICS-aligned industry that Trucost assigns natively (74 industries, full coverage). We also carry the full GICS hierarchy from Compustat and, for RQ5, the Sustainable Industry Classification System (SICS) from SASB. GICS and SICS are built on different principles, and that difference is what RQ5 exploits.

GICS, maintained jointly by MSCI and S&P Dow Jones Indices, classifies a firm by its principal business activity, determined primarily by the revenues each activity generates (MSCI and S&P Dow Jones Indices, 2023; Bhojraj et al., 2003). It is a market-oriented scheme: firms are grouped by the products and end markets their sales come from. Each firm receives a single code in a four-level hierarchy of 11 sectors, 25 industry groups, 74 industries, and 163 sub-industries. Because revenue is the assignment criterion, GICS aligns with exactly the quantity an input–output model scales. The EEIO estimate routes a firm's revenue through a GICS-defined sector intensity, so the modeled figure is close to a GICS-industry average.

SICS takes the opposite organizing principle. Developed by SASB and now maintained within the IFRS Foundation under the ISSB (Sustainability Accounting Standards Board, 2023; Khan et al., 2016), it groups firms by shared sustainability-related risks, resource intensity, and value-creation model rather than by revenue line. The scheme spans 11 sectors and 77 industries, each tied to a tailored sustainability-disclosure standard. Its reclassification can cut across GICS. Oil-and-gas producers (GICS Energy) and metals-and-mining firms (GICS Materials) fall together in a single SICS extractives sector, because they share resource-depletion and emissions profiles despite selling into different markets. Granularity is therefore close across the two schemes, 74 versus 77

industries. They partition on different axes, which lets a comparison isolate the grouping principle from the number of cells. Appendix Table 13 summarizes the construction differences.

4 Methods

4.1 Measuring agreement

For a firm-year with modeled emissions m and disclosed emissions d (both positive), the error is the log difference

$$e \;=\; \ln m - \ln d.$$

A positive e means the model overstates. We summarize the distribution of e with its mean (the bias), its mean absolute value (MAE), and its root mean square (RMSE), all in log units. An MAE of one corresponds to a typical discrepancy of a factor of $e \approx 2.7$. On levels we also report the median absolute percentage error (MdAPE). Rank agreement matters when the data are used to screen or rank firms rather than to value a single footprint. To capture it, we compute the Spearman correlation between modeled and disclosed values *within* sector–year cells, which removes the mechanical agreement that pooling across sectors and years would induce. Scale-dependent disagreement is shown with Bland–Altman plots, which set the log difference against the average of the two log measures.

We report these statistics for the full universe and for the emerging-technology subset, broken out by scope, by firm-size tercile, by reporting year, and by data-availability lag. Throughout we distinguish the upstream Scope-3 comparison, our cleanest test, from Scope 1 and 2, where the modeled field incorporates the disclosure.

4.2 What predicts the gap

To ask which firm and sector characteristics are associated with a larger gap, we regress the signed and the absolute log error on firm characteristics. The specification includes sector and year fixed effects, with standard errors clustered by firm. Continuous regressors are winsorized at the 1st and 99th percentiles *within the estimation sample*. Covariates are firm size, supply-chain breadth (cost of goods sold over revenue), capital and research intensity, leverage, and profitability. They also include the provider's own data-quality score, the share of the footprint that is disclosed, an emerging-technology indicator, and a region. We include the count of disclosed categories as a coverage control but do not interpret its coefficient, since the disclosed total is by construction the sum of those categories. These regressions are associational; we make no causal claim.

4.3 A hybrid estimator

We then ask whether a model that combines financial and sector signals with the EEIO subcomponents can estimate disclosed Scope 3 more accurately than the pure input–output figure. The target is the disclosed upstream total in levels, and every model fits it through a log link rather than on a log-transformed target. Modeling the level directly matters downstream: a per-category total is then formed by summing predicted levels, which avoids the retransformation bias that exponentiating and summing log-space predictions would introduce. To keep the target well defined, we restrict this experiment to near-complete disclosers, those with at least six of the eight upstream categories. Conditioning on completeness this way keeps it from contaminating the comparison. Features include firm financials, sector and region indicators, and the modeled Scope-1 and Scope-2 levels. They also include the EEIO Scope-3 upstream estimate and its revenue intensity,

the provider's data-quality score, and an environmental rating. Including the EEIO estimate as a feature lets the model learn corrections to the input–output figure rather than replace it.

We fit two estimators: an interpretable penalized Tweedie GLM (power 1.5, log link) and a gradient-boosted tree ensemble trained to the same Tweedie objective. Validation proceeds two ways. First, firm-grouped cross-validation places all of a firm's years in the same fold, so the model is never tested on a firm it has seen. A forward split then trains on years through 2022 and tests on 2023–24, matching the practical use of forecasting a current footprint. Predictions return in level space; we take logs only to score them, comparing each model's out-of-sample error against the EEIO estimate's error on the identical test firm-years, in natural-log units. Beyond accuracy, we examine where the gain concentrates by sector and which features carry it.

As an extension we model each disclosed upstream category separately and reassemble a predicted total by summing the predicted category levels a firm reports. This isolates which parts of the value chain are learnable from financial signals and tests whether a category-by-category approach improves on a direct model of the total.

We implement the estimators in scikit-learn and LightGBM with fixed random seeds. Categorical features are one-hot encoded and numeric features standardized; missing covariates are median-imputed with a missingness indicator rather than dropped, so the estimation sample is held fixed across models. The Tweedie GLM carries an L2 penalty with its log link and is fit by quasi-Newton iteration to convergence. Its boosted counterpart uses moderate regularization: a few hundred trees, a learning rate near 0.03, and subsampling of rows and features. Neither is heavily tuned, since the question is whether a generic hybrid improves on the input–output figure rather than which estimator wins a prediction contest. Because both models estimate the conditional level through a log link, no exponentiation step intervenes between prediction and summation, so the per-category total inherits no Jensen retransformation bias. A log-target alternative would not be so clean: a model unbiased in logs is biased on levels, and summing such retransformed predictions would not recover the disclosed total. Modeling levels directly is what lets the reassembled total compete with a direct model of the total.

4.4 Financial signals and their rationale

The hybrid model exists to correct a single, specific limitation of EEIO. Its input–output estimate uses one firm-level input, revenue, scaled by a sector-average intensity, so it is blind to any way a firm's production structure departs from its sector's average. Every financial signal we include is chosen to proxy exactly that departure. Each feature maps to a channel through which a firm's value chain can differ from the sector norm that EEIO imposes.

Size comes first. Revenue and real assets scale the level of activity, and therefore of emissions, and they anchor the prediction. Supply-chain breadth, measured as cost of goods sold over revenue, is the signal we expect to matter most. It captures how much of a firm's value is bought in rather than produced in house. That share is the direct determinant of upstream Scope 3, and the margin on which a revenue-only estimate strays furthest. Capital intensity, the ratio of capital expenditure to revenue, proxies the physical-asset and energy footprint that distinguishes asset-heavy from asset-light firms within a sector. Research intensity, R&D over revenue, marks the knowledge-intensive, asset-light business models whose emissions sit in different value-chain categories than their sector average implies. Leverage and operating profitability enter as business-model and financing controls; we hold weaker priors about their sign and include them so that the structural signals are read net of them.

The model also sees the firm's own modeled direct and energy emissions, the EEIO upstream estimate, and that estimate's revenue intensity. A firm's Scope 1 and Scope 2 levels co-move strongly

with its value-chain footprint, since both scale with the physical operations behind them. Including the EEIO estimate itself lets the model learn a correction to the incumbent number rather than rebuild it from scratch. Reporting controls round out the set: the provider's data-quality score, the disclosed share of the footprint, and an environmental-pillar rating. These absorb variation in measurement and disclosure behavior, so the financial signals are not confounded with how a firm reports.

4.5 Choice of estimators

The target shapes the choice of model. Disclosed Scope 3 is non-negative, strongly right-skewed, and spans several orders of magnitude, and its variance grows with its mean. The per-category targets add a further feature. Many firms report a clean zero for a given upstream category, so those distributions carry a point mass at zero alongside a continuous positive tail. Ordinary least squares on raw levels would let the largest emitters dominate the fit and would predict negative emissions for small firms. A log-transformed target sidesteps the skew but cannot represent the zeros, and it must be retransformed to recover a level, which reintroduces bias. Both estimators we use respect the structure of the target instead, and both predict on the level scale.

The first is a penalized Tweedie generalized linear model with a log link and variance power 1.5. With a power between one and two, the Tweedie family is the compound Poisson-gamma distribution. It is defined on the non-negative line, places positive probability on an exact zero, and has a continuous, right-skewed positive part. That is precisely the shape of disclosed Scope 3, in total and category by category. Its variance function makes the variance proportional to the mean raised to that power, the model's account of how dispersion grows with scale. A power of 1.5 sits between the Poisson and gamma cases and matches the over-dispersion emissions display. The log link makes the model multiplicative, so a covariate shifts the footprint by a proportion rather than a fixed quantity. That is how revenue, assets, and supply-chain breadth act on emissions. Because the model estimates the conditional mean on the level scale, its predictions are strictly positive and need no retransformation, and its coefficients read as semi-elasticities. An L2 penalty disciplines the many correlated inputs, the financial ratios, the sector and region indicators, and the EEIO estimate among them, so collinearity does not destabilize the fit. This is the interpretable anchor. It asks whether a parametric, log-linear correction to the input–output figure is enough.

The second is a gradient-boosted tree ensemble trained to the same Tweedie deviance, with the same power and log link. Holding the loss fixed is deliberate. Differences between the two models then come only from functional form, so the distance between them measures the value of relaxing linearity, not the effect of changing the objective. The ensemble keeps the non-negative, multiplicative structure of the GLM but drops its linearity. That lets it capture the interactions we expect on economic grounds, such as a supply-chain-breadth effect that differs across sectors, and the saturating relationships a single slope cannot. Trees also accommodate features on very different scales, tolerate skew without transformation, and absorb missing values directly, so the estimation sample stays fixed across the two models. Light regularization, a few hundred shallow trees with row and feature subsampling, guards against overfitting the heavy tail.

Together the pair brackets the interpretability-versus-flexibility trade-off on a common loss, which is what makes the comparison informative. When the flexible model improves on EEIO and the linear one does not, the useful correction is nonlinear rather than a simple reweighting of the inputs. Modeling levels through a log link, rather than modeling logs, is also what lets the per-category predictions be summed into a total without Jensen retransformation bias. We stop short of heavier machinery, deep networks or exhaustive tuning, on purpose. The question is whether a generic, well-specified hybrid improves on the input–output figure, not which estimator wins a

tuning contest. A transparent pair answers that more credibly than a black box would.

4.6 Comparing industry classifications (RQ5)

A final test asks whether the sector taxonomy itself drives the firm-level error. We attach the SASB Sustainable Industry Classification System (SICS) by bridging each firm's `gvkey` to ISIN through the Compustat security tables. The SICS master is then joined on ISIN. Coverage reaches 97% of the comparison firm-years. SICS resolves to 77 industries against GICS's 74, so the two schemes are compared at matched granularity on the firms classified under both.

We read the comparison three ways. First, we decompose the variance of log disclosed Scope-3 upstream, and of its revenue intensity, into between-sector and within-sector components. The decomposition uses a random-intercept model estimated by the one-way ANOVA method of moments. Its intraclass correlation, between over total variance, measures how much structure each taxonomy pushes between cells. Second, because SICS is designed to bite only where it regroups firms, we split the sample by whether a firm's SICS industry pools several GICS sectors. Within each split we compare within-cell intensity dispersion under the two schemes. Third, we re-compute the within-sector-year rank agreement of the modeled against the disclosed figure under SICS cells rather than GICS cells, which is the accuracy metric a screen would feel.

4.7 Financial materiality (RQ6)

A last question turns from the taxonomy to the issue. SASB's materiality map flags, for each SICS industry, whether GHG emissions are a financially material topic. We attach that flag through the same gvkey-to-SICS bridge, marking the 21 industries SASB designates as GHG-material. Materiality is therefore an industry-level treatment, so it cannot enter alongside an industry fixed effect, and every regression clusters its standard errors on the SICS industry.

Two outcomes follow the two research questions. For RQ-A we regress the signed and the absolute log error on the materiality flag, adding controls in steps. Those steps are disclosure completeness, the firm's emission intensity, then size with region and year fixed effects. The intensity control is the firm's EEIO-modeled Scope-1 (direct) intensity. Being high in heavy industries yet absent from the Scope-3 error, it separates a materiality effect from a carbon-intensity effect, where a Scope-3 intensity would be mechanically confounded. For RQ-B we regress a disclose-or-not indicator over the covered universe on the flag, with size and region and year fixed effects. Among disclosers we also regress the number of disclosed categories on the same terms.

We close with a placebo. The placebo repeats both tests for six other SASB issues, from Energy Management to Competitive Behaviour, ordered by how carbon-intensive their material industries are. If the materiality result is really an intensity effect, the raw pattern should track that ordering and should vanish under the intensity control for every issue, whatever the issue's subject.

4.8 Reproducibility

The full pipeline runs end to end from a single command, with random seeds fixed and every sample filter logged. It covers extraction, cleaning, sample construction, the accuracy and regression analyses, the hybrid models, and the robustness checks. Emission units, currency conversions, and identifier matches are recorded at each step.

5 Results

5.1 Accuracy by scope

Table 2 reports the headline comparison (RQ1). For direct and energy emissions the agreement was strong: the within sector–year rank correlation was 0.86 for Scope 1 and 0.93 for Scope 2, with small absolute errors. This is reassuring but partly mechanical, because Trucost adopts the disclosed Scope-1/2 figure when a firm reports one. The near-zero median percentage error for those scopes is the signature of that pass-through.

The upstream value chain was different. For Scope-3 upstream, where the modeled figure is an independent EEIO estimate, the rank correlation fell to 0.69 and the mean absolute log error rose to 1.66. That MAE is a median discrepancy of about a factor of five. Downstream Scope 3 sat in between, and the Scope-3 total was the noisiest. Figure 1 shows the agreement directly: the cloud of points is wide, and its spread depends on the magnitude of the estimate. Such a pattern marks a model that is close to right on average but unreliable firm by firm. A second view, Figure 2, shows that the model overstated relative to disclosures throughout, and that the overstatement grew along the value chain.

Part of that overstatement is not model error. Because the disclosed total sums only the categories a firm reports, an incompletely disclosing firm appears to emit far less than the model estimates for its full upstream chain. We quantify this below; it is the reason we treat disclosure completeness as a first-order feature of the data rather than a nuisance.

5.2 Heterogeneity in upstream accuracy

Accuracy was far from uniform (Table 3; RQ2). It improved sharply with firm size: among the largest tercile the upstream MAE was 1.48 and the pooled rank correlation 0.70, while among the smallest it deteriorated to 2.83 and 0.27. These are pooled within-cell correlations (ρ_{pool} in Table 3); pooling across sectors generally lifts the figure, so they are not on the same footing as the within-sector–year ρ_{avg} of Table 2. Input–output estimation applies sector-average intensities to revenue, so it is weakest where firm-specific structure departs most from the sector average. The clearest cases are small firms and, as the next subsection shows, sectors with distinctive supply chains. Across reporting years the gap was broadly stable. Stability also held across the data-availability lag: figures that took longer to appear were not systematically less accurate. That argues against stale data as an explanation for the divergence. In the emerging-technology subset the average accuracy was close to the universe (upstream MAE 1.61 against 1.66), but it varied widely within. Software and automobile components were estimated comparatively well, electronics-and-instruments and communications equipment comparatively poorly. Figure 3 displays the sector-level dispersion.

5.3 Drivers of the gap

Table 4 regresses the absolute gap on firm characteristics (RQ3). The strongest correlate was the coverage control, the count of disclosed categories. Its coefficient is large and precise, yet it is a near-identity: summing more categories makes the disclosed total mechanically closer to the modeled total. For that reason we do not read it as a behavioral driver. That dominance is itself the finding: the apparent modeled-versus-disclosed gap is governed first by how completely a firm discloses, and any accuracy statistic must be read conditional on that.

Holding coverage fixed, the firm fundamentals did far less work. Supply-chain breadth, the ratio of cost of goods sold to revenue, was associated with a larger gap (coefficient 0.44 on the absolute error, $t = 3.7$). Firms that buy in more of their value chain are exactly those for whom

a revenue-based estimate strays furthest. Once size is measured with revenue independent of the Trucost model, it is at most weakly associated with a smaller gap. The stronger size effect in the base specification is partly an artifact of a size proxy that shares inputs with the model. A higher data-quality score from the provider predicted a smaller gap, as one would hope. Neither the environmental rating nor the data-availability lag independently predicted the gap.

5.4 Hybrid-model performance

The input–output figure can be improved, though only with a flexible estimator (RQ4). On near-complete disclosers, a gradient-boosted model with a Tweedie (log-link) objective lowered out-of-sample RMSE against the disclosed value by 11.3% under firm-grouped cross-validation. Under a forward split that trains on the past and tests on the future, the reduction reached 14.5% (Table 5). An interpretable penalized Tweedie GLM did not share that gain. Its log-space error exceeded the EEIO estimate's, which tells us the useful correction is nonlinear rather than a simple reweighting of the inputs. Because both models estimate the disclosed level directly through a log link, their predictions carry no retransformation bias. Improvement was uneven across sectors. Gains concentrated where supply chains depart most from the revenue-intensity assumption, led by passenger airlines (71%), consumer-staples retail (54%), specialty retail (50%), and electric utilities (44%). In sectors that EEIO already handles well, the boosted model added little. Figure 4 shows it leaned on the modeled Scope-1 level, the EEIO estimate itself, and firm size and intensity. Those features let it learn sector-conditioned corrections to the input–output figure rather than discard it.

5.5 Per-category modeling and reassembly

Modeling each upstream category on its own (Table 6, Panel A) shows which parts of the value chain are learnable from financial signals. Fuel-and-energy and employee-commuting emissions were predicted best, with purchased goods close behind; upstream leased assets and the residual "other" bucket stayed nearly unpredictable. Summing the predicted category levels into a total beat a direct model of that total (Panel B). Reassembly improved on EEIO by 16.5%, against 13.1% for the direct model. This agrees with the cross-provider finding of Nguyen et al. (2023). The agreement depends on a modeling choice. Because each category is estimated as a level through a log link, its predictions can be summed without the retransformation bias that adding exponentiated log-space predictions would carry. A log-target version of the same models would reintroduce that bias and erase the advantage. Used this way, the category view maps where the signal lives and yields a competitive aggregate.

5.6 Within-sector dispersion of emission intensity

A simple variance decomposition shows why a sector-average estimate cannot be accurate firm by firm. We split the variation in disclosed Scope-3 upstream emission intensity into a between-sector component and a within-sector component, using a random-intercept model across GICS industries (Table 7). About half of it is within sector: the intraclass correlation is 0.49 for intensity and 0.49 after removing firm size, falling to 0.37 unconditionally.

That within-sector half is a ceiling on any revenue-times-sector-intensity estimator. EEIO assigns every firm in an industry the same intensity, so it captures the between-sector component by construction but is blind to the within-sector component. With roughly half the intensity variation living inside sectors, a sector-average number is structurally unable to track the firm-level differences that drive the modeled-versus-disclosed gap. The decomposition thus restates the

headline accuracy result as a property of the data rather than of one model: the firm-level signal EEIO misses is real and large.

5.7 Does the classification scheme matter? (RQ5)

The taxonomy is not the lever. That within-sector half, just documented under GICS, is the obvious place to look for improvement: a better partition would move variation from within cells to between them. Re-running the decomposition under SICS instead of GICS does not (Table 8). At matched granularity the between-sector share is the same under both schemes (intraclass correlation 0.49), so each taxonomy leaves about half the intensity dispersion inside its cells. A sustainability-native partition does not shrink the firm-level heterogeneity a sector-average estimate cannot capture.

The sharpest test is the one SICS should win. SICS regroups firms that GICS separates, most visibly by pooling oil-and-gas and mining into a single extractives sector, so any advantage should concentrate on those reclassified firms. It does not: on the firms whose SICS industry spans two or more GICS sectors, within-cell intensity dispersion is no lower under SICS than under GICS (a 0.4% difference). The small edge SICS shows appears instead among the firms it leaves essentially where GICS does. On the accuracy metric that screening feels, within-sector-year rank agreement between the modeled and disclosed figures, SICS is marginally worse than GICS (0.67 against 0.68).

We read this as a clean null with a constructive implication. The firm-level unreliability of EEIO Scope-3 is intrinsic to routing revenue through a sector-average intensity, not an artifact of the market-based taxonomy GICS supplies. Even the taxonomy built expressly around sustainability and resource intensity, and even on the firms it deliberately regroups, does not organize the disclosed data any better.

5.8 Does financial materiality explain accuracy or disclosure? (RQ6)

The taxonomy is one expert judgment about emissions; financial materiality is another. Prior work uses that judgment to predict stock returns and price informativeness (Khan et al., 2016; Grewal et al., 2021); we ask whether it predicts where the estimate itself can be trusted. SASB's materiality map flags 21 SICS industries for which GHG emissions are a financially material issue, the industries on which financed-emissions accounting, carbon-risk pricing, and climate disclosure rules concentrate. A natural worry is that the estimate is least reliable exactly there. The data say the opposite (Table 9, Panel A). In GHG-material industries the modeled figure is nearly unbiased, overstating disclosed upstream emissions by a mean of 0.12 in logs, against 0.81 in the rest. Its mean absolute error is lower, 1.43 against 1.72, and its rank agreement higher.

That reassurance is genuine, yet it is not a materiality effect. Panel B adds controls in steps. The raw materiality coefficient on the bias is -0.69, and it survives conditioning on disclosure completeness. It then collapses to -0.12 and loses significance the moment we condition on the firm's direct carbon intensity, and the same pattern holds for the absolute error. Material industries are estimated more accurately because they are carbon-intensive, the kind of large, process-emitting firm a revenue-based model fits well, not because their emissions are financially material. The relevant intensity is the firm's carbon intensity, not the financial intensity of materiality that Consolandi et al. (2022) relate to returns, namely a count of how many issues an industry deems material.

A placebo across other SASB issues settles the point (Table 10). We repeat the test for six further issues and order them by how carbon-intensive their material industries are. The raw accuracy advantage tracks that ordering, not the issue's subject. It is strongest for the dirtiest issue sets, Critical Incident Risk and GHG. The effect turns positive for Physical Impacts of Climate Change, whose material industries, real estate, insurance, and health care, are the cleanest in the

sample. Conditioning on direct intensity removes the effect for every issue but the sprawling 33-industry Energy Management set, which a single intensity summarises only loosely. No issue's materiality, GHG included, carries information about estimation accuracy beyond how carbon-intensive a firm is.

Materiality does not govern disclosure either (Panel C). Whether a firm reports upstream Scope 3 at all, and how many categories it reports, are unrelated to GHG materiality once size and region are fixed. The disclose-or-not coefficient is half a percentage point on a 7% base, statistically indistinguishable from zero, and the placebo returns the same null for all seven issues. Disclosure follows firm size and the regulatory region, not a financial-materiality label. That pattern matches the finding that firms disclose the categories easiest to compute rather than the most material (Nguyen et al., 2023; Klaaßen and Stoll, 2021).

Taken with the SICS null, the message is consistent. The firm-level structure that governs both the accuracy of estimated Scope-3 and the supply of disclosed Scope-3 lives in continuous fundamentals, emission intensity, size, and region. It does not live in the sustainability-native categories, taxonomy or materiality, built to capture it.

5.9 Robustness

The conclusions were stable. Winsorizing the error at the 1st/99th percentiles barely moved the upstream MAE (1.66 to 1.63). Restricting to firms with at least three years of disclosure left the large error intact (1.46), so it is not an artifact of first-time disclosers. Findings were insensitive to which borderline industries enter the emerging-technology definition. Adding a Compustat-versus-Worldscope source indicator, and its interaction with supply-chain breadth, left the breadth coefficient stable and significant, so the international gap-fill does not drive the result. Its own coefficient was positive, consistent with the harder-to-cover international firms carrying somewhat larger gaps. Because the log error is invariant to the intensity denominator, the accuracy conclusions are unchanged whether emissions are scaled by revenue, assets, or cost of goods sold. Full robustness tables appear in the appendix.

6 Discussion

The picture that emerges is consistent with what input–output practitioners have long suspected but rarely measured at this scale. EEIO Scope-3 estimates are reasonable in the aggregate and informative for ranking, but they are not firm-level measurements. For direct emissions the modeled and disclosed series agree, though largely because the provider adopts disclosures into the modeled field. Along the upstream value chain the estimate stands on its own, and there a typical firm's modeled and disclosed figures differ by about a factor of five. Rank agreement of the kind portfolio screening relies on is moderate at best. Divergence is widest exactly where the method's central assumption is least apt, namely that emissions scale with revenue at a sector-average intensity. Worst affected are small firms and firms that buy in more of their value chain.

Two implications follow for the users of these data. The first concerns measurement. Anyone who differences a modeled against a disclosed Scope-3 figure is, without care, measuring disclosure completeness as much as model error. Its modeled total covers the whole upstream chain, while the disclosed total covers only the categories a firm chose to report. Our evidence shows this completeness effect dominates the cross-sectional variation in the apparent gap. Studies and screens that compare "estimated versus reported" emissions should therefore condition on category coverage or restrict to complete reporters. A second implication concerns use. For the many applications that consume a single Scope-3 number (financed-emissions inventories, climate stress tests, target

baselines), the firm-level noise we document is inherited downstream. Treating EEIO Scope-3 as a rank signal rather than a level is the more defensible use. Better still, a flexible hybrid correction improves the level by 11 to 15% out of sample. That suggests it can be sharpened cheaply where it matters most, with data the user often already has.

For data providers, folding disclosures into the modeled field is sensible for delivering a best estimate. That practice still makes the product harder to evaluate, and it can mislead a researcher who assumes the "modeled" series is independent. Publishing the disclosed and modeled components separately, along with the category coverage behind a disclosed total, would let users judge reliability for themselves. The per-category results carry a methodological lesson. Building a total from category models is a viable route to a competitive aggregate. Its condition is that each category be modeled on its level, so the predictions sum without a retransformation bias. Modeled in logs and exponentiated, the same categories would compound errors and the reassembly would underperform. This category view therefore does double duty, showing which parts of the chain carry a learnable signal and supplying a total that rivals a direct model.

The classification scheme is not where the problem lives. We re-ran the entire comparison under the IFRS-endorsed sustainability taxonomy, SICS, which groups firms by resource intensity rather than revenue. It organized the disclosed data no better than GICS, even on the firms it deliberately regroups. The firm-level unreliability is therefore a property of revenue-based estimation itself, not of the market-oriented sector map EEIO happens to use. A reader inclined to dismiss the gap as a wrong-taxonomy artifact can set that explanation aside.

Several limitations bound these conclusions. Our reference is disclosure, not truth. Self-reported Scope 3 carries its own boundary and measurement error, so our statistics describe agreement with what firms report, and the residual divergence mixes model error with disclosure error. Because the disclosed upstream total is a constructed sum of reported categories, we lean on the completeness-conditioned and complete-reporter analyses. We study one provider's EEIO implementation. The qualitative pattern, accuracy strong for direct emissions but weak and structure-dependent for upstream Scope 3, should generalize to other revenue-based estimators, though the magnitudes need not. Recall too that the data-availability lag is a proxy for disclosure timing rather than a filing date. Our sample also predates mandatory Scope 3 reporting. The disclosure patterns we document, including the absence of a materiality effect, reflect a largely voluntary regime that the CSRD, ISSB S2, and the SEC climate rule may reshape. Finally, the driver analysis is associational: it identifies the characteristics that travel with a larger gap, not the causes of it.

7 Conclusion

Estimated Scope 3 emissions have become infrastructure for climate finance and policy, yet how well they match what firms disclose has gone largely unmeasured. Using a provider that records both a modeled and a disclosed value for the same firm-year, we benchmark input–output Scope-3 estimates against disclosures across a global panel. The modeled figures track direct emissions well in rank but diverge sharply for the upstream value chain, where they are genuinely independent of disclosure. That divergence is largest for small firms and supply-chain-heavy businesses. Before anything else, though, the apparent gap is a function of how completely a firm discloses. A flexible gradient-boosted model that corrects the input–output estimate with financial and sector signals cuts the out-of-sample error by 11 to 15 percent. Those gains land most heavily in the sectors the input–output method serves worst.

The contribution is less a verdict on input–output estimation than a map of its reliability. Where firms disclose, the disclosure should anchor the estimate. When they stay silent, EEIO remains a

defensible default, though its firm-level upstream figures are best used as ranks and, when a level is needed, are worth correcting. For data providers, separating the modeled and disclosed components and reporting category coverage would let users see the reliability we had to reconstruct. We see two natural extensions. One replaces disclosure with an independently audited reference where one exists, so that model error can be separated from disclosure error. Another would extend the hybrid correction to the downstream categories and to providers beyond the one studied here.

References

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

Berg, F., Kölbel, J. F., and Rigobon, R. (2022). Aggregate confusion: The divergence of ESG ratings. *Review of Finance*, 26(6):1315–1344.

Bhojraj, S., Lee, C. M. C., and Oler, D. K. (2003). What's my line? A comparison of industry classification schemes for capital market research. *Journal of Accounting Research*, 41(5):745–774.

Bolton, P. and Kacperczyk, M. (2021). Do investors care about carbon risk? *Journal of Financial Economics*, 142(2):517–549.

Bolton, P. and Kacperczyk, M. (2024). Are carbon emissions associated with stock returns? Comment. *Review of Finance*, 28(1):107–109.

Busch, T., Johnson, M., and Pioch, T. (2022). Corporate carbon performance data: Quo vadis? *Journal of Industrial Ecology*, 26(1):350–363.

Consolandi, C., Eccles, R. G., and Gabbi, G. (2022). How material is a material issue? stock returns and the financial relevance and financial intensity of ESG materiality. *Journal of Sustainable Finance & Investment*, 12(4):1045–1068.

Crawford, R. H., Bontinck, P.-A., Stephan, A., Wiedmann, T., and Yu, M. (2018). Hybrid life cycle inventory methods – a review. *Journal of Cleaner Production*, 172:1273–1288.

Davis, S., Dumit, A., Li, M., Maldonado, Y., Steffen, M., Stevenson, M., Boldyreva, T., and Suh, S. (2026). The importance of multiregional accounting for corporate carbon emissions. *Nature Communications*, 16.

Goldhammer, B., Busse, C., and Busch, T. (2017). Estimating corporate carbon footprints with externally available data. *Journal of Industrial Ecology*, 21(5):1165–1179.

Grewal, J., Hauptmann, C., and Serafeim, G. (2021). Material sustainability information and stock price informativeness. *Journal of Business Ethics*, 171(3):513–544.

Han, Y., Gopal, A., Ouyang, L., and Key, A. (2021). Estimation of corporate greenhouse gas emissions via machine learning. In *Tackling Climate Change with Machine Learning Workshop, ICML 2021*. arXiv:2109.04318.

Hertwich, E. G. and Wood, R. (2018). The growing importance of scope 3 greenhouse gas emissions from industry. *Environmental Research Letters*, 13(10):104013.

Kalesnik, V., Wilkens, M., and Zink, J. (2022). Green data or greenwashing? Do corporate carbon emissions data enable investors to mitigate climate change? *The Journal of Portfolio Management*, 48(10):119–147.

Khan, M., Serafeim, G., and Yoon, A. (2016). Corporate sustainability: First evidence on materiality. *The Accounting Review*, 91(6):1697–1724.

Kitzes, J. (2013). An introduction to environmentally-extended input-output analysis. *Resources*, 2(4):489–503.

Klaaßen, L. and Stoll, C. (2021). Harmonizing corporate carbon footprints. *Nature Communications*, 12:6149.

Leontief, W. (1970). Environmental repercussions and the economic structure: An input-output approach. *The Review of Economics and Statistics*, 52(3):262–271.

Li, X., Katafuchi, Y., Moran, D., Yamada, T., Fujii, H., and Kanemoto, K. (2024). Systematic underreporting in corporate Scope 3 disclosure. Research Square preprint. Working paper.

Matthews, H. S., Hendrickson, C. T., and Weber, C. L. (2008). The importance of carbon footprint estimation boundaries. *Environmental Science & Technology*, 42(16):5839–5842.

Minx, J. C., Wiedmann, T., Wood, R., Peters, G. P., Lenzen, M., Owen, A., et al. (2009). Input-output analysis and carbon footprinting: An overview of applications. *Economic Systems Research*, 21(3):187–216.

MSCI and S&P Dow Jones Indices (2023). GICS: Global industry classification standard methodology. MSCI Inc. and S&P Dow Jones Indices. Available at https://www.msci.com/our-solutions/indexes/gics.

Nguyen, Q., Diaz-Rainey, I., Kitto, A., McNeil, B. I., Pittman, N. A., and Zhang, R. (2023). Scope 3 emissions: Data quality and machine learning prediction accuracy. *PLOS Climate*, 2(11):e0000208.

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

Partnership for Carbon Accounting Financials (2020). The global GHG accounting and reporting standard for the financial industry. Technical report, PCAF.

Piñero, P., Heikkinen, M., Mäenpää, I., and Pongrácz, E. (2015). Sector aggregation bias in environmentally extended input output modeling of raw material flows in Finland. *Ecological Economics*, 119:217–229.

Science Based Targets initiative (2023). SBTi corporate net-zero standard. Technical report, Science Based Targets initiative.

Serafeim, G. and Velez Caicedo, G. (2022). Machine learning models for prediction of Scope 3 carbon emissions. Working Paper 22-080, Harvard Business School.

Steen-Olsen, K., Owen, A., Hertwich, E. G., and Lenzen, M. (2014). Effects of sector aggregation on CO2 multipliers in multiregional input-output analyses. *Economic Systems Research*, 26(3):284–302.

Su, B., Huang, H. C., Ang, B. W., and Zhou, P. (2010). Input-output analysis of CO2 emissions embodied in trade: The effects of sector aggregation. *Energy Economics*, 32(1):166–175.

Suh, S. and Huppes, G. (2005). Methods for life cycle inventory of a product. *Journal of Cleaner Production*, 13(7):687–697.

Sustainability Accounting Standards Board (2023). Sustainable industry classification system (SICS). IFRS Foundation. Available at https://sasb.ifrs.org/.

World Resources Institute and World Business Council for Sustainable Development (2011). Corporate value chain (Scope 3) accounting and reporting standard. Technical report, Greenhouse Gas Protocol, WRI and WBCSD.

Zhang, B., Caron, J., and Winchester, N. (2019). Sectoral aggregation error in the accounting of energy and emissions embodied in trade and consumption. *Journal of Industrial Ecology*, 23(2):402–411.

Table 1: Sample composition. The panel covers firms with a Trucost record over 2005–2024. Within it, the modeled-versus-disclosed comparison sample is the set of firm-years that carry both an EEIO-modeled and a company-disclosed Scope 3 upstream value (disclosed Scope 3 begins in 2020).

	Firm-years	Firms	Scope-3 upstream pairs
Emerging-technology subset	47,750	8,628	3,328
Other industries	219,179	36,232	13,116
Full universe	266,929	44,860	16,444

Notes. The emerging-technology subset is defined on GICS-aligned industry names (semiconductors and equipment, electrical equipment and storage, electric-vehicle and auto components, solar/wind, data-center software and internet, electronics hardware and communications). Financial covariates and GICS classifications are merged from Compustat (North America and Global) and, where Compustat is missing, Worldscope, via `gvkey`; financial coverage on the comparison sample is 99%.

Table 2: Accuracy of EEIO-modeled emissions against company-disclosed emissions, by scope (full universe). Error is the natural-log difference $e = \ln(\text{modeled}) - \ln(\text{disclosed})$. Bias is the mean signed error. MAE and RMSE are its absolute and root-mean-square; MdAPE is the median absolute percentage error on levels; ρ_{avg} is the Spearman rank correlation computed *within* each sector–year cell and then averaged across cells.

Scope	n pairs	Bias	MAE	RMSE	MdAPE	ρ_{avg}
Scope 1	29,544	0.59	0.62	1.54	0.00	0.86
Scope 2 (location)	29,318	0.26	0.29	0.97	0.00	0.93
Scope 3 upstream	16,444	0.65	1.66	2.45	0.74	0.69
Scope 3 downstream	11,566	0.96	1.02	2.54	0.00	0.85
Scope 3 total	17,047	1.27	1.81	3.02	0.62	0.73

Notes. Bias, MAE and RMSE are in natural-log units; an MAE of 1.66 corresponds to a typical multiplicative discrepancy of roughly a factor of five. MdAPE near zero for Scopes 1–2 reflects that Trucost adopts the disclosed figure as its reported value when a firm discloses. Those scopes are therefore not an independent test of the model. Scope 3 upstream, which is approximately pure revenue-intensity EEIO, is the cleaner benchmark. Disclosed Scope 3 totals are summed from the reported upstream (downstream) categories; see Section 4. The rank correlation ρ_{avg} is the equal-weighted mean of the within-cell Spearman correlations across sector–year cells, which strips out the cross-sector ordering a pooled correlation would reward; for upstream emissions it is 0.69, against a pooled full-sample value of 0.76. It is reported on a different basis from the pooled ρ_{pool} of Table 3 and is not directly comparable to it.

Table 3: Heterogeneity in Scope-3 upstream accuracy, by firm size, reporting year, and data-availability lag. Columns as in Table 2, except that ρ_{pool} is the Spearman correlation pooled across all firm-years in each row's cell.

	n	Bias	MAE	RMSE	ρ_{pool}
Panel A. Firm-size tercile (by real revenue)					
Large	10,867	0.47	1.48	2.20	0.70
Middle	4,418	0.78	1.82	2.60	0.53
Small	1,023	1.77	2.83	3.73	0.27
Panel B. Reporting year					
2020	2,020	1.07	1.86	2.66	0.71
2021	2,450	0.84	1.67	2.49	0.72
2022	2,125	0.29	1.44	2.06	0.75
2023	4,404	0.50	1.63	2.37	0.78
2024	5,445	0.67	1.70	2.54	0.79
Panel C. Data-availability lag tercile					
Short	5,779	0.62	1.62	2.41	0.78
Middle	5,036	0.73	1.70	2.47	0.73
Long	4,854	0.43	1.50	2.21	0.74

Notes. Size terciles use real, FX-converted revenue (independent of the Trucost model). The data-availability lag is the time from a firm's fiscal year-end to the earliest Trucost extraction of the figure (median ≈1.2 years). It proxies disclosure timing but also reflects the data provider's processing cadence. Accuracy is roughly flat across the lag terciles. We report the pooled ρ_{pool} here rather than the cell-averaged ρ_{avg} of Table 2, because the finer sector–year cells within a size tercile are too thin to support a per-cell rank correlation: in the smallest tercile about half of these cells hold fewer than three firm-years and are undefined, and roughly four in five hold fewer than ten. Because ρ_{pool} pools across sectors within each row, it also reflects cross-sector ranking and typically exceeds the cell-averaged ρ_{avg}; the two are not directly comparable.

Table 4: What predicts the size of the modeled-versus-disclosed gap? Coefficients from regressions of the *absolute* log error in Scope-3 upstream on firm and sector characteristics, with sector and year fixed effects and standard errors clustered by firm. Continuous regressors are winsorized at the 1st/99th percentiles within the estimation sample. This analysis is associational, not causal.

	(1) Base	(2) + Financials
Supply-chain breadth (COGS/revenue)		0.437^{***}
		(0.118)
Log revenue (size)[a]	-0.064^{***}	-0.087^{**}
	(0.012)	(0.043)
Log assets (size, real USD)		0.034
		(0.042)
Capital-expenditure intensity		−0.056
		(0.169)
Leverage		0.049
		(0.112)
Operating ROA		−0.069
		(0.292)
Trucost data-quality score	-0.051^{**}	-0.053^{**}
	(0.025)	(0.026)
GHG disclosure (%)	-0.0055^{***}	-0.0054^{***}
	(0.0006)	(0.0007)
Data-availability lag (years)	0.004	
	(0.039)	
Emerging-technology	−0.074	−0.143
	(0.147)	(0.150)
Environmental-pillar score		−0.001
		(0.001)
Disclosure breadth (coverage control)	-0.471^{***}	-0.430^{***}
	(0.013)	(0.014)
Sector fixed effects	Yes	Yes
Year fixed effects	Yes	Yes
Firm-years	15,500	13,197

Notes. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$. Standard errors in parentheses, clustered by firm. [a]Column (1) uses a Trucost-implied revenue measure; column (2) uses real, FX-converted Compustat/Worldscope revenue that is independent of the Trucost model. Disclosure breadth (the count of reported Scope-3 upstream categories) is included as a coverage control. Because the disclosed total is by construction the sum of those categories, its coefficient is a mechanical identity and is not interpreted as a behavioral driver.

Table 5: Out-of-sample accuracy of the hybrid models against disclosed Scope-3 upstream, relative to the EEIO estimate. Both models estimate the disclosed level through a log link, so their predictions need no retransformation. Errors are root-mean-square (RMSE) and mean-absolute (MAE) in natural-log units. "vs. EEIO" is the change in RMSE against the EEIO baseline on the identical test firm-years, where a negative value is an improvement. The sample is gated to near-complete disclosers (≥ 6 of 8 upstream categories; $n = 4{,}940$).

	Firm-grouped CV			Forward split (2023–24)		
Model	RMSE	MAE	vs. EEIO	RMSE	MAE	vs. EEIO
EEIO baseline	1.280	0.931	—	1.244	0.952	—
Tweedie GLM	1.559	1.138	+21.9%	1.771	1.353	+42.4%
LightGBM (Tweedie)	1.135	0.781	−11.3%	1.063	0.756	−14.5%

Notes. Firm-grouped cross-validation places all of a firm's years in the same fold to prevent leakage; the forward split trains on ≤2022 and tests on 2023–24 (n = 1,907). Both models use a Tweedie variance power of 1.5 with a log link, fit on the raw disclosed level: the GLM as an L2-penalized regression, LightGBM as a boosted ensemble. The interpretable GLM does not beat EEIO in log space, so the useful correction is nonlinear. Features are firm financials, sector and region indicators, the modeled Scope-1/2 levels, the EEIO Scope-3 upstream estimate and its intensity, the Trucost data-quality score, and the S&P environmental-pillar score.

Table 6: Per-category Scope-3 upstream modeling. Panel A reports out-of-sample accuracy of a separate model for each disclosed upstream category. Panel B compares the reassembled per-category total against a direct model of the total and against the EEIO estimate, on the same gated sample as Table 5.

Panel A. Accuracy by upstream category (Tweedie LightGBM, grouped CV)			
Category	Disclosers	RMSE	R^2
Cat. 3, fuel and energy	12,225	1.58	0.71
Cat. 7, employee commuting	12,338	1.57	0.51
Cat. 1, purchased goods	13,326	2.42	0.49
Cat. 5, waste generated	12,658	2.23	0.44
Cat. 2, capital goods	9,310	1.99	0.44
Cat. 4&6, transport and travel	7,073	2.23	0.40
Cat. 8, upstream leased assets	3,085	2.49	0.19
Other upstream	287	3.74	0.09
Panel B. Reassembly versus a direct total model			
Approach		RMSE	vs. EEIO
EEIO baseline		1.280	—
Direct total model		1.112	−13.1%
Per-category reassembly		1.068	−16.5%

Notes. R^2 is out-of-sample (firm-grouped) and computed in log space. Each category is modeled on its raw level with a tweedie-objective LightGBM. The reassembled total sums the predicted category levels over the set a firm disclosed, and is compared to its disclosed total on that same set. Summing levels avoids the retransformation bias that adding exponentiated log-space predictions would carry, which is why the reassembly beats the direct total model rather than trailing it. Categories follow the GHG Protocol upstream set (categories 1–8); categories 4 (upstream transportation) and 6 (business travel) are reported jointly by the provider and modeled together.

Table 7: Within-sector dispersion of disclosed Scope-3 upstream emissions (GICS industries). A random-intercept decomposition splits the (log) variance into between-sector and within-sector components; the intraclass correlation (ICC) is the between-sector share. The intensity row scales emissions by revenue.

Specification	Between-sector	Within-sector	ICC
Unconditional	1.62	2.80	0.37
Size-adjusted	1.12	1.16	0.49
Intensity (emissions/revenue)	1.12	1.17	0.49

Notes. Variance components are in natural-log units, estimated by the one-way ANOVA method of moments across the 72 GICS industries present in the gated sample ($n = 4{,}951$; 4,900 for the size-adjusted and intensity rows, which require revenue). About half the intensity variation is within sector, the part a sector-average EEIO estimate cannot capture.

Table 8: Does the classification scheme matter? GICS versus SICS at matched granularity (RQ5). Panel A decomposes the variance of disclosed Scope-3 upstream emission intensity into between- and within-sector components (the intraclass correlation, ICC) on the firm-years classified under both schemes. Panel B reports within-cell intensity variance for firms SICS leaves with their GICS peers ("aligned") versus firms whose SICS industry pools two or more GICS sectors ("regrouped"). Panel C reports the within sector–year rank agreement of the modeled against the disclosed figure on the full comparison sample.

	GICS	SICS
Panel A. Variance decomposition, intensity ($n = 4{,}821$)		
Industry cells	72	76
Between-sector ICC, unconditional	0.367	0.343
Between-sector ICC, intensity	0.492	0.496
Within-cell intensity variance	1.149	1.144
Panel B. Within-cell intensity variance, by regrouping		
Aligned firms ($n = 615$)	0.886	0.811
Regrouped, ≥ 2 GICS sectors ($n = 4{,}192$)	1.169	1.173
Panel C. Rank agreement, modeled vs disclosed ($n = 15{,}880$)		
Mean within sector–year Spearman	0.684	0.672

Notes. Emission intensity is disclosed Scope-3 upstream over revenue, in logs. The ICC is the between-sector share of variance from a random-intercept model (ANOVA method of moments); a higher ICC means the taxonomy pushes more structure between cells. SICS is bridged to the panel via `gvkey`→ISIN (Compustat security tables) and covers 97% of the comparison firm-years. Across every panel the two schemes are within noise, and SICS shows no advantage on the firms it regroups, so the choice of taxonomy does not explain the firm-level gap.

Table 9: Does financial materiality explain accuracy or disclosure? (RQ6). Panel A compares modeled-versus-disclosed Scope-3 upstream accuracy for firms whose SICS industry SASB flags as GHG-material versus the rest, on the comparison sample. Panel B reports the GHG-material coefficient on the (log) gap as controls are added; the signed error is the bias (positive means the model overstates) and the absolute error is the dispersion. Panel C reports the GHG-material coefficient on disclosure. Materiality varies at the SICS-industry level, so standard errors cluster on the SICS industry; intensity is the firm's EEIO-modeled Scope-1 (direct) intensity, high in heavy industries yet not a component of the Scope-3 error.

Panel A. Accuracy by GHG materiality (Scope-3 upstream)

	Material	Non-material
Firm-years	4,016	11,864
Bias (mean signed log error)	0.118	0.807
MAE (mean absolute log error)	1.434	1.722
Within sector–year rank (Spearman)	0.704	0.658

Panel B. GHG-material coefficient on the gap (t in parentheses)

	Bias	Absolute error
(1) raw	−0.690 (−3.1)	−0.288 (−2.0)
(2) + completeness	−0.640 (−3.1)	−0.252 (−2.9)
(3) + direct (Scope-1) intensity	−0.119 (−0.6)	−0.073 (−0.7)
(4) + size, region/year FE	−0.086 (−0.5)	−0.025 (−0.2)

Panel C. GHG-material coefficient on disclosure

	Coefficient	Baseline
Discloses upstream (size, FE)	+0.005 (1.0)	0.071
Categories | disclose (size, FE)	+0.057 (0.4)	4.28

Notes. Panel B regresses the signed and absolute log error on the GHG-material indicator and the listed controls (n = 15,690 in the full specification). Panel C regresses a disclose-or-not indicator over the Trucost-covered universe (n = 222,919) and, among disclosers, the number of reported upstream categories (n = 15,814). The accuracy advantage of material industries is large in the raw comparison but vanishes once the firm's direct carbon intensity is held fixed, so it reflects emission intensity, not financial materiality. Materiality does not predict disclosure.

Table 10: Placebo across SASB materiality issues (RQ6). For each SASB general-issue category we re-run the accuracy and disclosure tests with that issue's material-industry flag in place of GHG. "Dirtiness" is the gap in EEIO-modeled Scope-1 (direct) emission intensity between the issue's material and non-material industries, in logs. "Bias, raw" and "Bias, +intensity" are the material coefficient on the signed Scope-3 upstream error before and after conditioning on Scope-1 intensity (and disclosure completeness); "Disclosure" is the material coefficient on the probability of disclosing. Rows are ordered by dirtiness.

SASB issue	Dirtiness	Bias, raw	Bias, +intensity	Disclosure	Industries
GHG Emissions	+3.44	−0.69 (−3.1)	−0.12 (−0.6)	+0.005 (1.0)	21
Critical Incident Risk	+3.34	−0.79 (−3.4)	−0.33 (−1.5)	+0.003 (0.5)	16
Legal & Regulatory Env.	+2.54	−0.61 (−1.8)	+0.05 (0.3)	−0.002 (−0.2)	5
Ecological Impacts	+2.16	−0.22 (−1.1)	+0.03 (0.2)	−0.004 (−0.9)	14
Energy Management	+0.63	−0.57 (−2.4)	−0.35 (−2.2)	+0.001 (0.3)	33
Competitive Behaviour	−0.29	−0.21 (−1.0)	−0.06 (−0.5)	+0.003 (0.7)	11
Physical Impacts of Climate	−1.38	+0.38 (0.6)	−0.23 (−0.4)	+0.007 (0.8)	8

Notes. t-statistics (SICS-industry-clustered) in parentheses; "Industries" is the number of material SICS industries. The raw bias effect tracks dirtiness rather than the issue's subject: it is largest for the most carbon-intensive issue sets and turns positive for Physical Impacts, whose material industries (real estate, insurance, health care) are the cleanest. Conditioning on direct carbon intensity removes the effect for every issue except Energy Management, whose broad 33-industry set is imperfectly summarised by a single intensity. Disclosure incidence is insignificant for all seven issues.

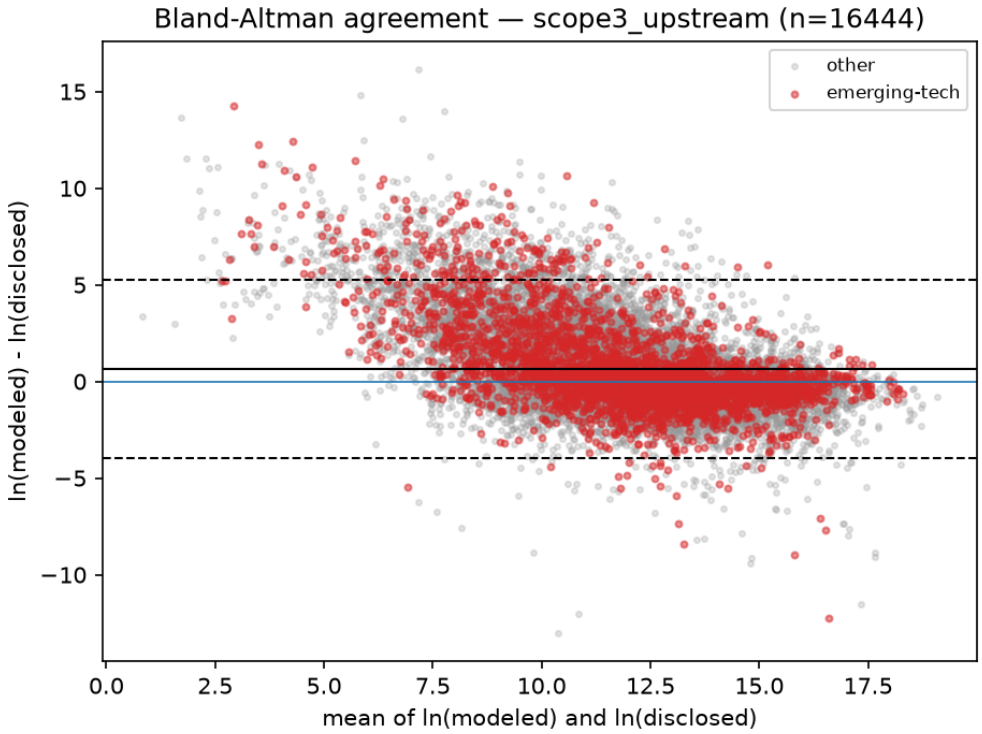

Figure 1: Agreement between EEIO-modeled and disclosed Scope-3 upstream emissions (Bland–Altman). The horizontal axis is the average of the two log measures and the vertical axis their difference, ln(modeled) − ln(disclosed). A solid line marks the mean bias and the dashed lines the 95% limits of agreement; points in emerging-technology industries are highlighted. That wide, scale-dependent spread is the visual counterpart of the large Scope-3 upstream error in Table 2.

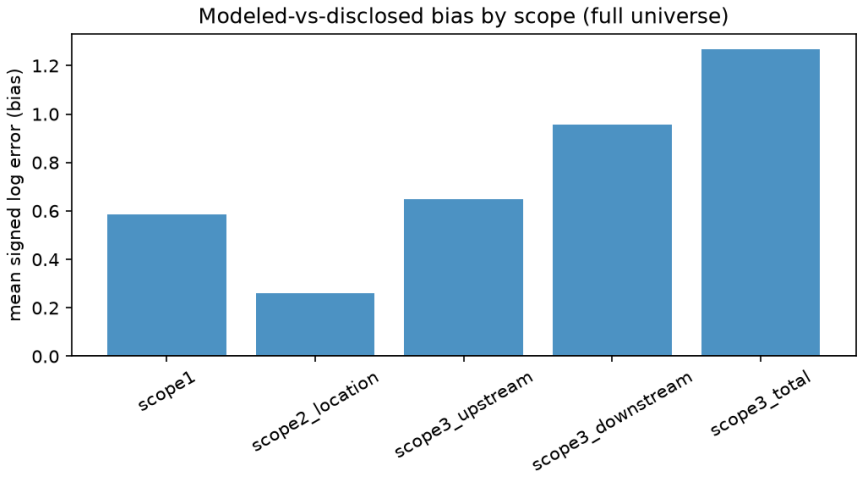

Figure 2: Mean signed log error (bias) by scope in the full universe. The model overstates relative to disclosures throughout, and the bias grows from direct emissions to the upstream and downstream value chain. That rise is in part a mechanical consequence of comparing a coverage-complete modeled total against a sum of only the categories a firm discloses.

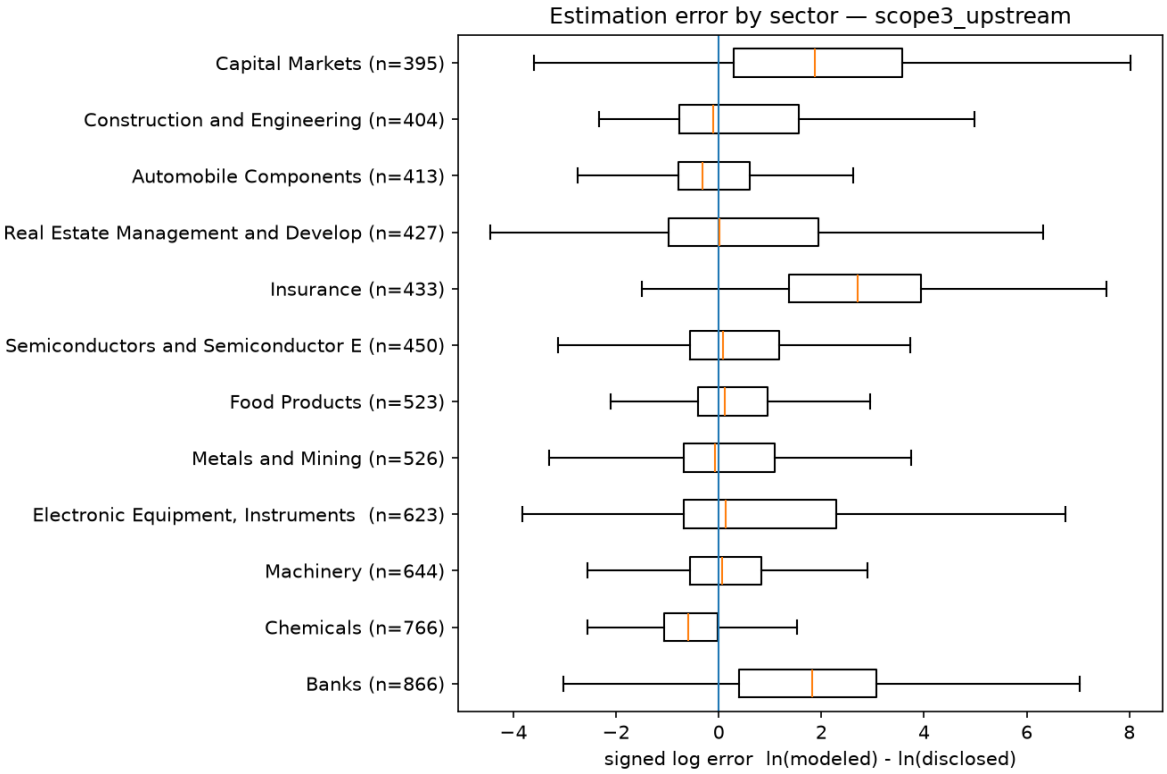

Figure 3: Distribution of the signed Scope-3 upstream log error across the sectors with the most observations. Boxes show the interquartile range and median; the vertical line marks zero error. Both the central tendency and the dispersion of the modeled-versus-disclosed gap vary substantially across sectors.

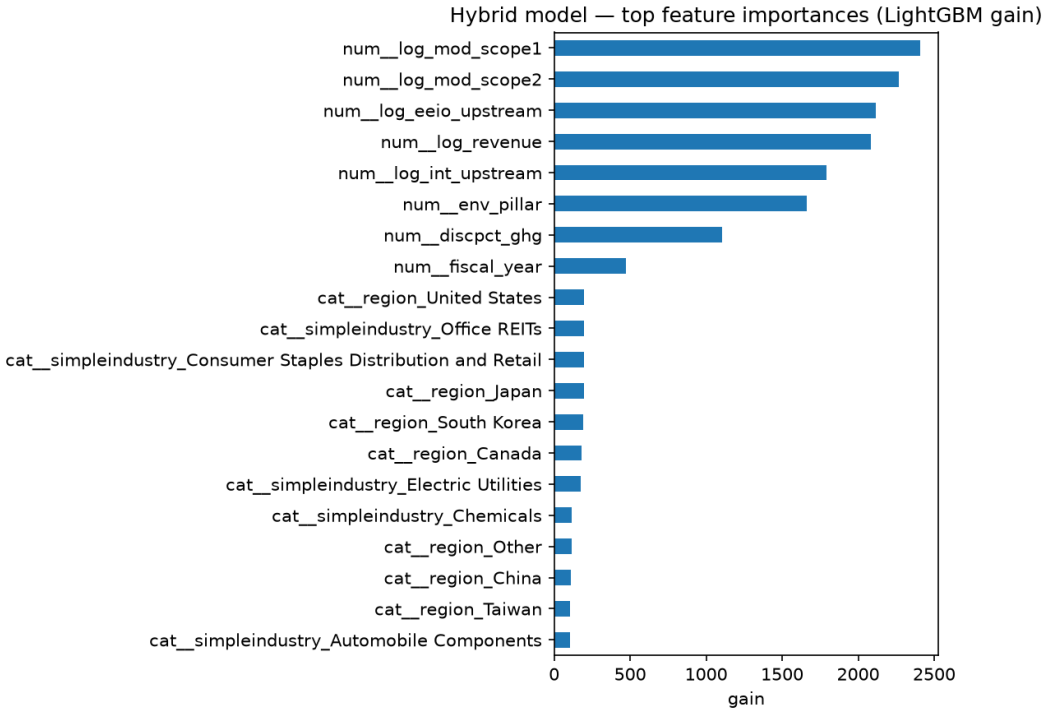

Figure 4: Feature importances (gradient-boosting gain) for the hybrid model of disclosed Scope-3 upstream. The modeled Scope-1 level, the EEIO Scope-3 estimate itself, and firm size and revenue intensity carry most of the signal. These features let the hybrid learn sector-conditioned corrections to the input–output estimate rather than discarding it.

A Variable construction and additional results

A.1 Constructed variables

Disclosed Scope-3 totals and completeness. The provider reports no single "upstream total as reported." We sum the eight reported upstream categories (GHG Protocol categories 1–8 plus an "other upstream" bucket) and record the count of non-missing categories as the disclosure-completeness measure. A value of exactly zero is treated as missing.

Financial covariates and the Worldscope fill. Financials come from Compustat North America and Global, unioned by `gvkey`. Because the disclosing universe is about four-fifths non-U.S., we fill the remaining gaps from Worldscope, linking `gvkey` to ISIN through the Compustat security tables and ISIN to the Worldscope code. The fill is applied at the firm-year level so that all financial fields for a given firm-year come from a single source; a `source` flag records which. Worldscope "U.S. dollar" items are used for dollar levels, and currency-invariant ratios are used otherwise. Where the two sources overlap, cost-of-goods-to-revenue agrees with correlation 0.75 and a median absolute difference of 0.008.

Real, model-independent size. To avoid a size control that shares inputs with the Trucost model, the financial specifications use real revenue and assets converted to dollars. Local-currency Compustat Global and Worldscope values are converted with annual average cross-rates, rather than read off a Trucost-implied revenue.

Data-availability lag. The earliest Trucost extraction date for a firm-year minus its fiscal year-end, in years, winsorized at the 1st/99th percentiles. It proxies disclosure timing but also reflects the provider's processing cadence.

A.2 Financial-source robustness

Table 11 confirms that the international gap-fill does not drive the supply-chain-breadth result. Adding a Worldscope-source indicator, and its interaction with cost-of-goods-to-revenue, leaves the breadth coefficient stable and significant; the indicator is positive, consistent with harder-to-cover international firm-years carrying somewhat larger gaps.

Table 11: Financial-source robustness (dependent variable: absolute Scope-3 upstream log error; sector and year fixed effects; firm-clustered standard errors).

	(1) Financial	(2) + source flag	(3) + source × breadth
Supply-chain breadth (COGS/revenue)	0.437^{***}	0.470^{***}	0.541^{***}
Worldscope source indicator		0.413^{***}	0.732^{***}
Breadth × Worldscope			-0.732^{**}
Firm-years	13,197	13,197	13,197

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$, two-sided with firm-clustered standard errors. The supply-chain-breadth coefficient is stable and significant across the three specifications, so the Compustat-versus-Worldscope gap-fill does not drive it.

A.3 Robustness summary

Table 12 reports the Scope-3 upstream mean absolute error under alternative samples and treatments. The large error is stable.

Table 12: Robustness of the Scope-3 upstream mean absolute log error.

Specification	MAE
Baseline (full comparison sample)	1.66
Winsorized at 1st/99th percentiles	1.63
Firms with ≥ 3 years of disclosure	1.46
Emerging-technology, core definition	1.61
Emerging-technology, with borderline sectors	1.48

A.4 Reproducibility

The analysis runs end to end from a single command over the WRDS extracts, with fixed random seeds and logged sample filters. Proprietary data are not redistributed; the code and the variable dictionary reproduce every table and figure from the source databases.

A.5 GICS versus SICS: construction methodology

Table 13 contrasts the two industry classifications the paper uses. The schemes are close in granularity but partition firms on different axes: GICS on the revenue an input–output model scales, SICS on the resource intensity the model assumes constant within a sector.

Table 13: Construction methodology: GICS versus SICS.

	GICS	**SICS**
Maintained by	MSCI and S&P Dow Jones Indices (since 1999)	SASB, now within the IFRS Foundation / ISSB
Organizing principle	Principal business activity: products and end markets	Shared sustainability risks, resource intensity, business model
Assignment criterion	Revenues generated by each activity	Common sustainability profile, which can cut across product markets
Orientation	Market / revenue	Sustainability / resource use
Hierarchy	Sector, industry group, industry, sub-industry	Sector, industry (each mapped to a SASB standard)
Levels (counts)	11 / 25 / 74 / 163	11 / 77
Illustrative divergence	Oil & gas in Energy; metals & mining in Materials (separate sectors)	Both in one Extractives & Minerals Processing sector
Primary purpose	Investment analysis, portfolio and peer construction	Materiality-based sustainability disclosure
Role in this paper	`simpleindustry` (Trucost-native, 74 industries, 100% coverage); full GICS also from Compustat	SASB SICS mapping, used for the RQ5 comparison

Notes. Level counts for GICS are the post-2023 revision. GICS groups firms by the revenue dimension the EEIO model uses to scale sector intensities. SICS instead groups by the resource-intensity dimension the model holds constant within a sector. The two schemes therefore differ along the axis that governs firm-level EEIO error. Sources: MSCI and S&P Dow Jones Indices (2023); Bhojraj et al. (2003) for GICS; Sustainability Accounting Standards Board (2023); Khan et al. (2016) for SICS.