Data quality in the spotlight: a Hybrid-LCA approach to evaluating reported corporate carbon footprints

Author 1

ABSTRACT

Global warming is a substantial issue, and addressing it requires robust data on corporate environmental performance. However, there is still an information gap in indirect emissions — particularly Scope 3 emissions — due to the small number of reporting companies, greenwashing practices, and methodological differences across reports. In this study, we propose a new Hybrid-LCA approach for estimating corporate Scope 3 emissions for a large universe of companies following the Greenhouse Gas Protocol framework. Our approach recovers expected trends in emissions while maintaining the comparability and transparency needed for decision-making. Moreover, we observe an asymmetric bias with respect to self-reported data from the Carbon Disclosure Project. Our analysis indicates that this bias is partially related to poor environmental management practices in firm disclosures, which we proxy by factoring in the verification status of firms' reports. This work calls for greater scrutiny in Scope 3 reporting to ensure the accuracy of emission data and promote better environmental management practices.

Introduction

Climate change is one of the most pressing issues that humanity is facing today. The Intergovernmental Panel on Climate Change (IPCC) report[1] states that global surface temperatures have risen between 0.95-1.20°C in the past century, and most of it can be attributed to anthropogenic emissions. As a result, the past decade has been warmer than any multi-century period in the last 125,000 years. Immediate action is needed to reduce greenhouse gas (GHG) emissions to prevent further catastrophic impacts from global warming. However, according to the Climate Action Tracker[2], global warming will likely reach 2.7°C by the end of the century under current policies. Worryingly enough, recent studies show evidence of multiple climate tipping points at the risk of being triggered if temperatures rise above 1.5°C[3]. The consequences could include sea level rise from collapsing ice sheets, dieback of the Amazon rainforest and warm-water corals, and carbon release from thawing permafrost.

All organizations must reduce their direct GHG emissions to address the climate change challenge. But, direct emissions from industry are only 14% of the total when accounting for indirect emissions[4]. We also need to look into total GHG footprints so that companies cannot greenwash their emissions by offsetting their polluting activities to providers in other geographies[5]. Indirect emissions are also a concern for businesses and investors, given the climate risk they pose for the former and the transition risk transmitted to the latter[6].

The Green House Gas (GHG) Protocol's definition of carbon footprint[7]—the primary global standard for GHG emissions measuring—separates direct emissions from the organization (Scope 1), indirect emissions due to the consumption of electricity (Scope 2), and the rest of indirect emissions (Scope 3)[8]. Scope 3 emissions follow a 17-category taxonomy distinguishing between upstream and downstream categories (see table 1). But there are three main issues to overcome to have complete and quality Scope 3 information: the accessibility, the trustworthiness, and the comparability of data across organizations.

The first issue we encounter is the need for more companies to disclose Scope 3 emissions. There are different data providers gathering reports (Clarity AI, Bloomberg, MSCI, S&P (TruCost), Moodys, MorningStar (Sustainalytics), and ESG Book are some of the names). But Carbon Disclosure Project (CDP)—a non-profit organization that receives volunteer reports through a yearly sustainability questionnaire—is among the most influential in the industry. While previous studies have extensively used CDP[9–11], it only collected emissions for 2413 companies in its 2021 questionnaire.

The second issue is the trustworthiness of corporate reporting. The CDP questionnaire follows the GHG protocol guidelines in Scope 3 calculation, but the recommendations are not binding, and the accounting is done by each firm individually. Moreover, the reports are not necessarily audited by a third party—only 38.5% did in the 2021 CDP questionnaire—so their validity is at least questionable. As companies disclose their environmental performance, there are few incentives or resources to reveal the correct information, and therefore companies can incur under-reporting practices. Corporate declarations about environmental performance leadership have seen a sharp increase in recent years, as has the amount of literature pointing to greenwashing practices[12]. There is evidence that US consumer goods companies tend to under-report greenhouse gas emissions[13] and that the top environmental performers are the ones disclosing better environmental information[14]. Recent contributions also show that some firms adhering to sustainable initiatives (GRI, UNGC) fail to improve environmental impact substantively after doing so[15].

The third issue is related to the heterogeneity of carbon footprint estimation methodologies. The underlying methodological

Category	Sub-category	Includes
Scope 1	Single	Direct sources due to the operations of the organization
Scope 2	Single	Indirect sources related to the consumption of energy
Scope 3	Upstream	Indirect sources related to: - Purchased goods and services - Capital goods - Fuel and energy-related activities not included in scope 1 or 2 - Upstream transportation and distribution - Waste generated in operations - Business travel - Employee commuting - Upstream leased assets - Other upstream
	Downstream	Indirect sources related to: - Downstream transportation and distribution - Processing of sold products - Use of sold products - End-of-life of sold products - Downstream leased assets - Franchises - Investments - Other downstream

Table 1. Outline of the GHG Protocol taxonomy of organizations' emissions.

diversity makes comparisons across companies hard, spoiling the results' implications. Life Cycle Analysis (LCA) methodologies — evaluation of inputs, outputs, and their environmental impact according to the ISO 14040 standard[16] — are the most common technique to estimate Scope 3 emissions[17]. And there are two main approaches to LCA modeling: Process-based LCA and Environmentally extended input-output (EEIO) LCA.

On the one hand, process-based life cycle assessment (LCA) is a bottom-up approach that involves breaking down a production system into a series of processes representing the product's life cycle[18]. This approach can obtain accurate results while following the GHG Protocol to estimate all the categories. However, upstream, downstream, or sideways truncations[19] can reach errors between 50%[20] and 87%[21]. Minor deviations in the employed emission factors result in significant errors[22]and all of this assuming cost and time constraints are not an issue[23]. On the other hand, EEIO-LCA is a top-down approach that examines macroeconomic trade patterns and the emissions associated with them[24–26]. Through this global approach, we avoid truncation errors by default. However, EEIO-LCA does not follow the GHG Protocol. Not only because it does not split emissions into Scope 1, 2, and 3 categories but because it can exclude crucial aspects such as final use[27], end-of-life management[28], and capital goods[29].

Several studies have combined Process-based-LCA and EEIO-LCA techniques to counteract the limitations of the two, giving rise to Hybrid-LCA methodologies[30]. However, Hybrid-LCA models are typically case-specific, focusing on specific processes or industries, so comparability is still an issue. Moreover, they usually do not follow the GHG Protocol separation of Scope 1, 2, and 3. Furthermore, the high specificity combined with the need for more tools and software available to support practitioners makes Hybrid-LCA methods hard to implement in the industry. Furthermore, if not implemented correctly, Hybrid-LCA does not necessarily yield more accurate results than Process-based LCA[31]. In summary, LCA and EEIO incur different errors and biases. And state-of-the-art methodologies (Hybrid-LCA) are hard to implement at scale and only sometimes follow the GHG protocol.

As a result of the accessibility, trustworthiness, and comparability issues, the literature casts severe doubts around the validity of Scope 3 data[9,22,32,33]. In fact, a recent study has found that corporate reports in the tech sector omit half of the total Scope 3 emissions[34]. But the need for a Scope 3 dataset meeting the quality requirements is even more significant than the challenges in building it. Under an Environmental, Social, and Governance (ESG) framework, investors and companies could leverage the information to protect against transition and climate risks (financial materiality) while contributing to solving global warming (stakeholder materiality)[35]. There are already studies linking ESG to financial performance[36], and while there are still negative results[37,38], it has been shown that noise in ESG data could be behind them[39]. Moreover, incoming

policies such as the EU Taxonomy or the SFDR regulation are forcing investors to declare the sustainability of their investments. Providing a transparent methodology to estimate Scope 3 emissions (following the GHG protocol taxonomy) of a large universe of companies while assuring the comparability of the estimates across companies is critical. Recent literature efforts have estimated Scope 3 emissions for a large universe of companies using machine learning[40]. However, corporate reports are the source of truth for this modeling approach, so it potentially inherits any biases already present in the data.

The contributions of this work are two-fold. On the one hand, we propose a new Hybrid-LCA method to tackle the abovementioned Scope 3 data issues. We solve the accessibility gap through a scalable methodology that follows the GHG emissions protocol. The transparency of the methodology provides trustworthiness, and because we use the same model across companies, we solve the comparability issue. On the other hand, we compare our independent Scope 3 estimates to CDP self-reported Scope 3 emissions. Although we recover general trends in the data, we observe a significant level of heterogeneity in the residuals, which suggests a misalignment between our model and CDP self-reported data. The residuals are not symmetric, pointing to a possible under-reporting bias for some companies. We also find that companies verifying their Scope 3 reports with a third-party provider report closer to our model, underlining the low trustworthiness of some of the CDP data points. Overall, our findings highlight the importance of using a consistent and reproducible methodology to estimate Scope 3 emissions accurately and identify potential biases in self-reported data.

The rest of this paper is structured as follows: in the next section, we discuss the model's results and the differences between the model and reported data, and later provide a discussion on the implications of our results. The mathematical formulation and implementation of the Hybrid-LCA model, as well as the CDP data used to compare it, are described in the methods section,

Results

The proposed Hybrid-LCA model allows us to estimate Scope 1, 2 and 2 emissions at the geography and industry level, assigning an emission intensity (TnCO2/MUSD) to the company based on its revenue distribution across geographies and industries, as well as the number of employees. This way we independently estimated Scope 3 emissions for 30k public companies, out of which 2431 reported to the CDP yearly questionnaire in 2021.

The results of the Hybrid-LCA model can be analyzed at the macro (industry and country) and micro (company) levels. On the one hand, we present the industry and country-level trends, and validate their agreement with previous literature. On the other hand, we compare the Scope 3 estimates at the company level to CDP self-reported data. Finally, we analyse the relationship between the residuals and third party verification of the report.

Industry and country level

First, we can understand the macro trends the model is leveraging by examining the industry and country emission intensity vectors (first term of equation 10).

In Figure 1 we show the median intensities by industry and emission type. In concordance with previous literature[4], Scope 1 and 2 emissions are the minor shares of emissions for most of the industries. "Mining and quarrying of energy producing products" has the most significant overall GHG footprint; high Scope 1 emissions due to the heavy machinery used and a sizeable downstream impact due to the use of its products for energy generation. Scope 1 emissions in the utilities and transportation sectors account for most of the overall GHG emissions. The most extreme case is "Air transportation", for which Scope 3 emissions accounted for only 36.8% of the total emissions. Overall, we found that tertiary sectors were less GHG intensive than secondary sectors, which are less intensive than primary sectors.

Focusing on the geographic side, we also observe some heterogeneity—although not as much as in the industry segmentation (Figure 1). As previously found in the literature[41], developed nations seem to have lower Scope 3 intensity of emissions (Figure 2). This result is aligned with the Environmental Kuznets Curve hypothesis[42,43], as well as from developed countries offsetting polluting activities to developing economies (carbon leakage hypothesis)[44–46].

Comparison to CDP self-reported data

Using our Hybrid-LCA model, we estimate the Scope 3 emissions intensities (TnCO2/MUSD) for 30k public companies. As explained in the methods section, we do that through company revenues' industrial and geographic footprint ($\vec{r}$) and the number of employees (n). Moreover, we compare the estimates with the information reported by companies to the CDP questionnaire 2021 (mainly corresponding to the fiscal year 2020).

We report the similarity between CDP and our estimates in Table 2. We observe an overall Spearman correlation of 0.689. This correlation is driven mainly by a 0.745 Spearman correlation of industry medians, with most industries (84.7%) in the range of the reported values. The share of industries for which we obtain a positive and significant Spearman correlation only reaches 8%—close to random if we consider the 5% significance level.

To further understand the discrepancies between the Hybrid-LCA and self-reported data, we visually compare the residuals of regressing CDP data against the Hybrid-LCA model for upstream (left) and downstream (right) in Figure 4 emissions. We

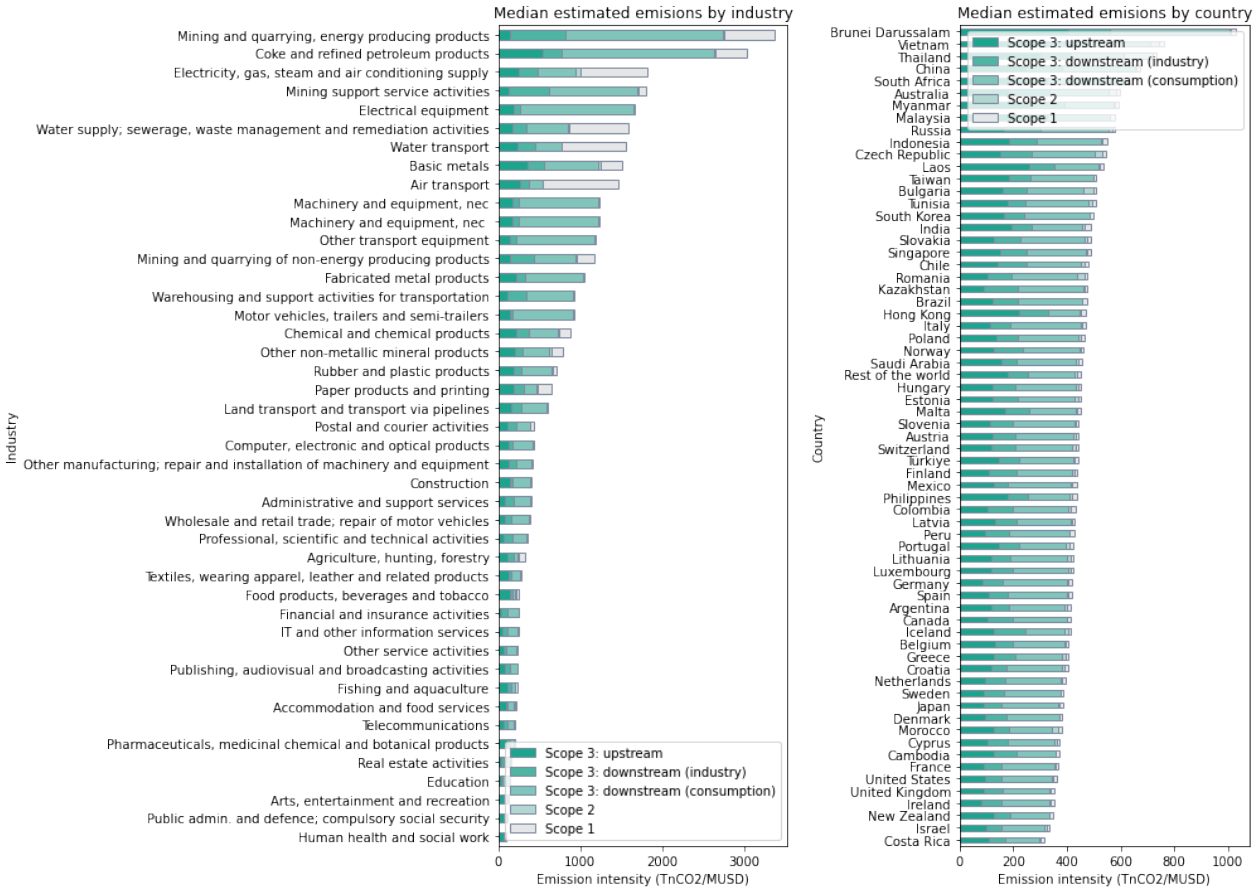

Figure 1. Median estimated intensity of emissions (CO2Tn/10^6USD) by industry (left) and country (right). We observe that the energy sector leads the footprint ranking, with the "Mining and quarrying, energy-producing products" at the top. This result is expected since mining leverages heavy machinery with many Scope 1 emissions. Also, the Use of sold products category of Scope 3 is very intensive—burning fossil fuels to produce electricity. We observe that the heterogeneity is smaller in the geographic case, where Scope 3 emissions account for most of the emissions across countries. Most developed economies can be found at the bottom of the graph, while the most intensive ones are the least developed. We explore this trend further in Figure 2.

Decision-maker requirements	Goodness-of-fit measure of Hybrid-LCA model relative to CDP self-reported data	Total	Upstream	Downstream
Overall company ordering	Overall spearman correlation	0.689	0.572	0.688
Industry ordering	Spearman correlation of industry-wise medians	0.745	0.636	0.725
Ordering of companies within an industry	Share of industries with a positive and significant within-industry spearman correlation	8.0%	11%	3.7%
Correct order of magnitude	Share of industry medians in the reported range	84.7%	92.6%	66.9%

Table 2. Goodness-of-fit of Hybrid-LCA model relative CDP self-reported data. We provide different error metrics related to decision-makers requirements. We observe that the Hybrid-LCA model recovers a substantial amount of the signal in CDP data, however it fails to reproduce most of the within-industry rankings.

observe two patterns: on the one hand, there is more significant heterogeneity in CDP data. On the other hand, we observe an asymmetric distribution of discrepancies between the two. While most of the datapoint fall around the identity line, there is a higher density of data points where the company reported fewer emissions than the Hybrid-LCA model. Moreover, this result seems independent of the source of Scope 3 emissions (upstream or downstream).

We observe several common phenomena between the residuals (reported vs estimated intensity) of the upstream and downstream components of Scope 3 emissions (Figure 4). While the residuals are approximately centered around 0, there is a second negative mode. As a result, the distribution skews are negative (-0.246 and -0.306 for the upstream and downstream components respectively, having removed the 1 and 99 percentiles as outliers). Moreover, these effects are more acute for the

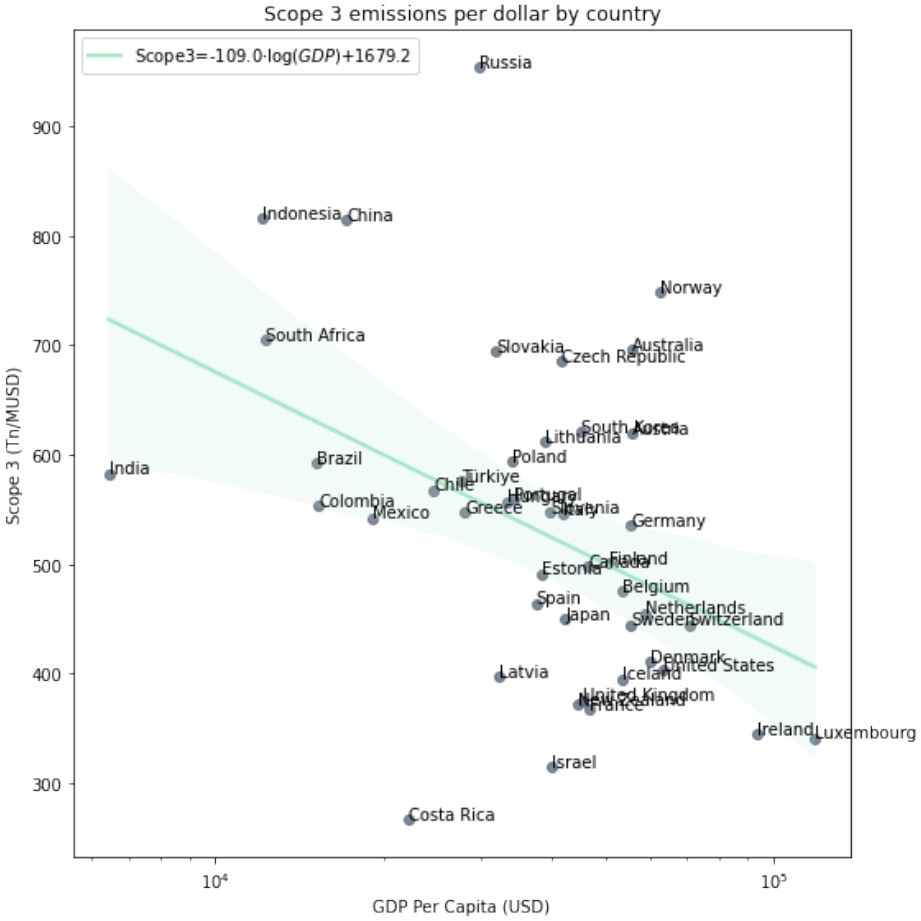

Figure 2. Total Scope 3 emission intensity (Tn/MUSD) by country vs. GDP Per Capita (USD). We estimate the total emissions by country by multiplying the intensities at the industry and country levels by their economic output, to then sum all the emissions by country and divide by the total output. We fit the Ordinary Least Squares (OLS) regression model $Scope3 = \alpha \cdot \log(GDP) + \beta$, the light area marks the 95% confidence interval for the regression estimate. A negative relationship is detected, aligned with the Environmental Kuznets Curve hypothesis and with developed countries offsetting polluting activities to developing economies.

subset of companies that hired a third-party provider to verify the reports. To further investigate this point we estimate the following two regression analysis:

$$\log(\Phi^i_{III}) = \beta_i \cdot \log(\hat{\Phi}^i_{III}) + \gamma_i \cdot Verification + \alpha_i \tag{1}$$

where $i \in (Upstream, Downstream)$, Φ^i_{III} is the CDP reported data, $\hat{\Phi}^i_{III}$ is the Hybrid-LCA estimated data, and $Verification = 1$ if the organization verified its emissions report with a third party provider, else $Verification = 0$.

We quantitatively compare the two sources via the regression model results (Table 3). First, the slopes β_i are both close to identity, revealing the alignment of the two signals. Second, the negative intercepts shows that CDP self-reported data stands below the Hybrid-LCA model. The effect is greater in the case of downstream emissions, but it is statistically significant in both cases. Finally, the verification dummy is positive and significant in both cases, signaling that firms verifying their Scope 3 report with a third-party provider provide 32.3% (and 28.7%) higher upstream (downstream) emissions, and are therefore closer to the Hybrid-LCA model. These results are consistent with the hypothesis that CDP Scope 3 data suffer from under-reporting bias.

Discussion

High-quality Scope 3 data is crucial for society to solve global warming, for investors to protect against transition risks, and for companies to avoid climate risks. In this work, we have identified the need for a Scope 3 emissions data set that solves the three

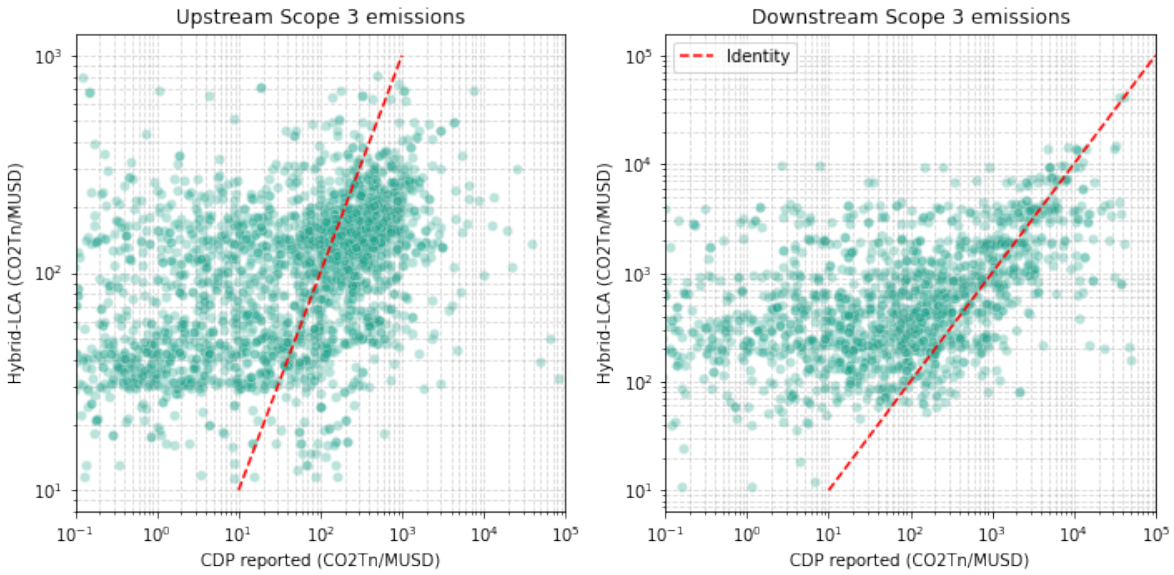

Figure 3. Comparison of company-level Scope 3 emisions, estimated via the Hybrid-LCA model vs self-reported to the CDP questionnaire.

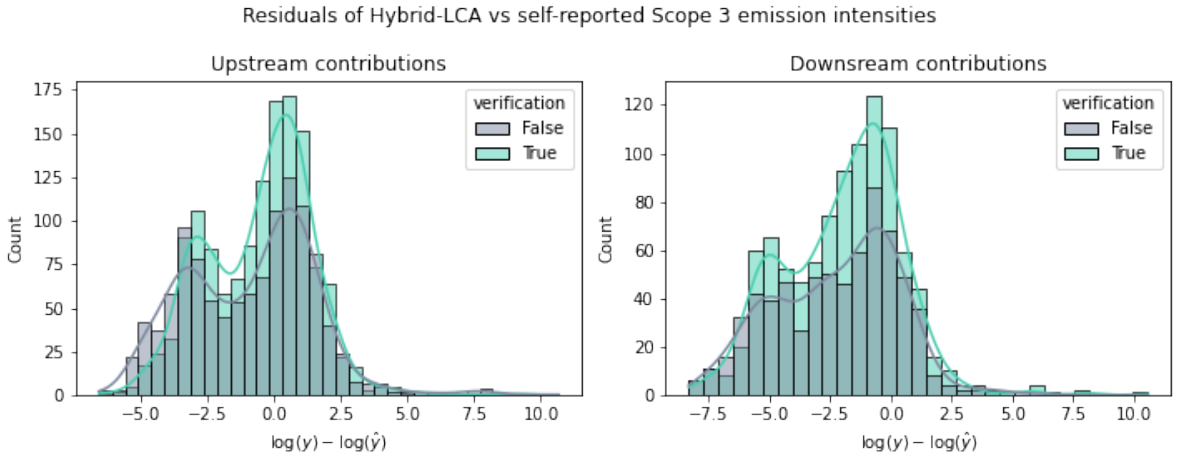

Figure 4. Residuals for the regression between self-reported and Hybrid-LCA modeled Scope 3 emissions for the upstream categories combined. We observe an asymmetric distribution with a long left tail and a second mode in the negative values. Data coming from 3rd party verified reporting companies seems to have a smaller left tail.

quality criteria: accessibility, trustworthiness, and comparability. Our Hybrid-LCA model is easily scalable to thousands of companies, while following the GHG protocol taxonomy of emissions. We hope to build trust through transparency in the methodology, and since we leverage the same model for all companies, we solve comparability.

At the macro level, the proposed model recovers meaningful trends in the emissions estimates. As expected, Scope 1 and 2 emissions are generally a minor share of emissions for most industries. The mining and quarrying of energy-producing products had the highest total emissions, while air transportation had the lowest share of Scope 3 emissions. Tertiary sectors were generally less intensive than secondary sectors, which were, in turn, less intensive than primary sectors. The results showed some heterogeneity when considering geographical location, although less than the industry breakdown. Moreover, as expected from the Environmental Kuznets Curve hypothesis and the offsetting of polluting activities to developing economies, developed nations displayed lower Scope 3 emission intensities.

The geographic and industrial activities of firms allow us to provide independent estimates of corporate Scope 3 emissions. And by leveraging the 2021 Carbon Disclosure Project questionnaire, we compare our model to self-disclosed corporate data.

	Upstream - Baseline	Upstream - Verification Dummy	Downstream - Baseline	Downstream - Verification Dummy
Dep. Variable	$\log(\Phi^U_{III})$	$\log(\Phi^U_{III})$	$\log(\Phi^D_{III})$	$\log(\Phi^D_{III})$
Estimator	PooledOLS	PooledOLS	PooledOLS	PooledOLS
No. Observations	2431	2431	1595	1595
Cov. Est.	Unadjusted	Unadjusted	Unadjusted	Unadjusted
R-squared	0.1343	0.1389	0.2055	0.2081
R-Squared (Within)	0.0000	0.0000	0.0000	0.0000
R-Squared (Between)	0.1354	0.1389	0.2056	0.2076
R-Squared (Overall)	0.1343	0.1389	0.2055	0.2081
F-statistic	376.72	195.84	411.93	209.18
P-value (F-stat)	0.0000	0.0000	0.0000	0.0000
$\log(\hat{\Phi}^U_{III})$	1.0015***	1.0151***		
$\log(\hat{\Phi}^D_{III})$			1.0165***	1.0231***
α	-0.7128***	-0.9521***	-2.1859***	-2.3959***
γ		0.3233***		0.2872**

Table 3. Estimates for the linear regression model $\log(\Phi^i_{III}) = \beta_i \cdot \log(\hat{\Phi}^i_{III}) + \gamma_i \cdot Verification + \alpha_i$, where $i \in (U, D)$, Φ^i_{III} are the CDP self-reported Scope 3 intensities, $\hat{\Phi}^i_{III}$ are the Hybrid-LCA estimates, and the *Verification* dummy captures when a third-party provider verified the self-report.

To perform the comparison, we defined a set of metrics that the decision-makers care about: the ability to rank companies, the ability to rank industries, getting the correct orders of magnitude, and being able to rank companies against their peers in the industry. While we recover most of the geographical and industrial trends, recovering within industry ordering of emissions intensities is a more challenging task. Future research could focus on providing confidence intervals around the company-reported data to understand the magnitude of the errors compared to the differences between industry peers.

Further analysis of the residuals between the proposed model and CDP self-reported data yields intriguing results. We observe an overall bias that makes the errors asymmetric (CDP emissions being generally lower). A possible explanation is a phenomenon of under-reporting, which we try to validate through an additional signal, with 38.5% of the firms in the comparison universe verifying their Scope 3 emissions report with a third-party provider. In this way, we draw two conclusions from our regression analysis. First, we observe that CDP reports are systematically below the Scope 3 estimates from our model. Second, we observe that the verified companies report significantly higher emissions, supporting the under-reporting hypothesis.

The results of this study have several important policy implications for decision-makers, regulators, and companies. For decision-makers, it is essential to use data estimated with the same model to ensure comparability. The model should also cover as many companies as possible to provide a comprehensive view of the portfolio. Additionally, the precautionary principle[47] suggests that decision-makers should use worst-case scenario assumptions when evaluating a company's emissions performance, as companies may have low incentives to report when they are underperforming. Regulators must provide tools for leveraging hybrid-LCA methods to estimate emissions accurately. Additionally, there should be enforcement of correct reporting to ensure that companies accurately disclose their emissions. Regulations such as the EU's SFDR and taxonomy are good examples of contributions in this direction. Furthermore, companies must report their emissions to provide transparency and accountability, even if they are bad performers. It is also crucial for companies to acknowledge their limitations and seek help when they cannot accurately compute their emissions. By following these guidelines, decision-makers, regulators, and companies can work together to reduce GHG emissions and mitigate the impacts of climate change.

Methods

Hybrid-LCA model

This section outlines the proposed model for estimating corporate Scope 3 emissions. In summary, our proposal is to modify classical Environmentally Extended Input Output (EEIO) models to follow the Green House Gas Protocol and assign intensities of emissions to companies based on their revenue footprint by industry and geography.

EEIO models take two aspects into account: (i)the global macroeconomic flows to map supply chain interdependencies and (ii) environmental stressors to convert the interdependencies into environmental footprints. In particular, EEIO models begin from the total sales volume or revenue (number of dollars) of goods and services going from supplier industry i to customer industry j, represented by matrix $Z = [z_{ij}]$. We compute the technical coefficient matrix as the share of output ($\vec{x}$) from industry i required to produce one dollar of output from industry j ($A = [a_{ij}] = [z_{ij}/x_j]$). We can express the total output that each industry must produce as the sum of its contributions to final demand, its direct industrial customers, second-tier industrial

customers (the customer's customers), and so on:

$$\vec{x} = (\mathbb{I} + A + AA + AAA + \cdots) \cdot \vec{f} = (\mathbb{I} - A)^{-1} \cdot \vec{f} = \mathscr{L} \cdot \vec{f} \tag{2}$$

where $\vec{x}$ is the vector of outputs per industry, $\vec{f}$ is the vector of final consumption, and $\mathscr{L}$ is the Leontief inverse matrix[24]. Given a unit of final demand, $\mathscr{L}$ provides the economic output each industry and country have to produce to satisfy it. Therefore, classical EEIO models derive the indirect upstream impact of an industry by multiplying by the emission factors (Tonnes of CO2 per dollar, also named emission intensities) of those industries satisfying its demand for goods and services[25]:

$$\vec{\Phi}^{U}_{indirect} = \mathscr{L}^{\top} \cdot \vec{\Phi}_{direct} \tag{3}$$

Where $\vec{\Phi}_{direct}$ is the vector of direct emissions per unit of output of the providers, and $\vec{\Phi}^{U}_{indirect}$ is the total upstream footprint.

However, three primary conceptual mismatches exist between the previous expression and the Scope 3 definition from the GHG protocol.

The first one is the double-counting of Scope 1 and 2 emissions. The identity matrix in Equation 2 stands for the industry's direct consumption and thus adds Scope 1 emissions to the total footprint. Scope 2 emissions—upstream footprint related to direct electricity providers—should also be removed. We do this by calculating the first-tier emissions related to electricity generation:

$$\vec{\Phi}_{II} = \left[\vec{u}^{\top} \odot \vec{\Phi}_{I}^{\top}\right] \cdot A \tag{4}$$

where $\odot$ stands for the Hadamard (or element-wise) product, and $u_i = 1$ only if i is an electricity production industry; otherwise, $u_i = 0$. The second conceptual mismatch is that the model completely misses the downstream emissions. As previously pointed out, the responsibility for all the downstream emissions must be shared between the producer and the consumer[48,49]. The income-based formalism of Input Output theory has been proposed as a solution[50,51], although only applied to national footprint calculations. Based on Ghosh's formulation of Input-Output theory[52], the downstream footprint is expressed as:

$$\vec{\Phi}^{D}_{indirect} = \mathscr{G} \cdot \vec{\Phi}_{direct} \tag{5}$$

where $\mathscr{G} = (1 - B)^{-1}$ is the Ghosh inverse matrix, where $B = [b_{ij}] = [z_{ij}/x_i]$ is closely related A, only in this case dividing by the inputs (see details in the supplementary materials). Provided one unit of input, $\mathscr{G}$ provides the economic output each industry and country will produce by consuming it.

Finally, the third mismatch is that some Scope 3 categories—such as "Employees commuting" and "Use of sold products"—are not covered by EEIO methodologies since they lay outside the supply chain.

We propose a new Hybrid-LCA model that fixes the three above-mentioned issues. We estimate emission intensities—tonnes of CO2 emissions by million US dollars—of Scope 3 emissions ($\vec{\Phi}_{III}$) by industry and geography adjusting EEIO models to the GHG protocol. To solve the double-counting issue, we remove Scope 1 ($\vec{\Phi}_{I}$) and Scope 2 ($\vec{\Phi}_{II}$) emissions from the upstream footprint to obtain upstream Scope 3 emissions ($\vec{\Phi}^{U}_{III}$):

$$\vec{\Phi}_{II} = \left[\vec{u}^{\top} \odot \vec{\Phi}_{I}^{\top}\right] \cdot A \tag{6}$$

$$\vec{\Phi}^{U}_{III} = \vec{\Phi}_{I}^{\top} \cdot [\mathscr{L} - \mathbb{I}] - \vec{\Phi}_{II} \tag{7}$$

We also remove Scope 1 emissions from the downstream footprint to obtain downstream Scope 3 emissions by industry and geography ($\vec{\Phi}^{D}_{III}$):

$$\vec{\Phi}^{D}_{III} = \vec{\Phi}_{I}^{\top} \cdot \left[\mathscr{G}^{\top} - \mathbb{I}\right] \tag{8}$$

We then take into account the emissions due to the use of sold products ($\vec{\Phi}^{C}_{III}$) that the company contributed to:

$$\vec{\Phi}^{C}_{III} = [\vec{\Phi}_{c}^{\top} \odot \vec{\gamma}] \cdot \mathscr{G}^{\top} \tag{9}$$

where $\vec{\gamma}$ provides the share of goods and services produced that are sold to the final consumer, and $\vec{\Phi}_{c}$ is the intensity of emissions of consumption. We additionally estimate the emissions from employees as $\vec{\Phi}^{E}_{III} = \vec{\Phi}^{T}_{e} \cdot n$, where $\vec{\Phi}^{T}_{e}$ are the emissions per employee and dollar, and n are the number of employees. To provide company-level Scope 3 estimates, we map industry

and geography-level intensities to companies through the geographic and industrial activities of the company expressed as percentages of the revenue in each of them ($\vec{r}$):

$$\hat{\Phi}_{III} = \left[\vec{\Phi}^{U}_{III} + \vec{\Phi}^{D}_{III} + \vec{\Phi}^{C}_{III} + \vec{\Phi}^{E}_{III}\right] \cdot \vec{r} \tag{10}$$

where $\hat{\Phi}_{III}$ are the company's total estimated Scope 3 emissions. One potential criticism of the model is that we are effectively applying a single emission factor by industry and country (except for the Employee commuting contribution, which depends on the number of employees). However, we consider this a feature since it allows us to preserve comparability and transparency—all the companies within the same exact industry and country are provided the same emissions. Moreover, public companies reporting to CDP have unique distributions of revenue per industry and geography, so the model provides unique emissions for each company.

To implement the methodology, we leverage a combination of data sets; The Input-Output tables come from the 2021 edition of OECD Inter-Country Input-Output (ICIO) Tables, covering 45 unique industries based on ISIC Revision 4 and 66 countries for the years 1995 to 2018. For the geographic and industrial revenue shares, we leverage Factset's RBICS and GEOREV data sets, which provide company revenue shares by sector and geography respectively. Since the Input-Output output tables and the RBICS database use different industrial classifications, we manually mapped the RBCIS business lines to the ISIC Revision 4 to use a common classification.

We derive the aggregate Scope 1 geography and industry emission intensities $\vec{\Phi}_I$ leveraging the Clarity AI emissions dataset (and assuming that public companies are representative of the whole industry). We assign 32578 available Scope 1 company emissions for the fiscal year 2020 to their respective countries and industries through the geographic and industrial revenue shares dataset. Clarity AI has such extensive coverage thanks to various machine learning algorithms through which they efficiently collect, estimate and check data. To ensure that the assigned companies are representative of each country-industry, we only consider companies having at least 90% of their revenue in that industry in particular. After removing outliers based on the percentiles 7.5 to 92.5 limit, we compute the average intensity of the companies belonging to that industry and that country weighted by their market share. We follow the same procedure to estimate $\vec{\Phi}_c$ and $\vec{\Phi}_e$, only this time aggregating the emission intensities reported by companies in the "Use of sold products" and "Employees commuting" categories in the CDP questionnaire. And last, we use the number of employees from Clarity AI's dataset.

Carbon Disclosure Project self-reported data

As previously mentioned, another source of Scope 3 emissions is the Carbon Disclosure Project (CDP) self-reported data. CDP issues a yearly questionnaire to capture most of the relevant ESG metrics for organizations. Most of the 2020 Scope 3 emissions were reported to the 2021 questionnaire, which contained the following question: Account for your organization's gross global Scope 3 emissions, disclosing and explaining any exclusions. The reporting company disclosed emissions at the category level (i.e., Tonnes of Scope 3 emissions for each of the 17 categories). Also, they could provide additional information: evaluation status of the category and methodology used, among others.

This part of the questionnaire was answered by 2413 companies in 2021, although not all organizations report all 17 categories. In Figure 5, we show that most of the categories are not considered relevant by most companies. And even if they are, a significant share of companies do not estimate them. A good example is the Use of sold products category, which is only reported by 26%—not what we expected since almost every consumption has unavoidable emissions associated.

Some companies verified their answers in the questionnaire with a 3rd party provider. As of December 08, 2022, CDP acknowledged six organizations as accredited verification solutions providers. These organizations are independent of the organizations that have gathered and provided the data and those that will use the data. They are accredited to perform certification under internationally-recognized standards, including relevant ISO or ISAE standards (CDP accepts 43 accepted verification standards). In the 2021 questionnaire, 38.5% of companies confirmed having verified their climate information disclosure.

Comparison of Hybrid-LCA and self-reported data

The discrepancies between the Hybrid-LCA model and CDP data could be due to different reasons. The obvious one is the methodology heterogeneity, given that all estimates are produced with different models. But overlooking this point, there are two further possible explanations. On the one hand, the proposed Hybrid-LCA model assigns the same Scope 3 emission intensities to companies fully operating in the same sectors and geographies. On the other hand, CDP data lacks comparability and transparency, not to mention the lack of incentives to report correctly, so there is room for underreporting practices.

To compare the two approaches we define four metrics to compare our estimates to self-reports, capturing relevant information decision-makers care for. The first and most essential desiderata is to provide the correct emissions intensity ranking, which we proxy with the correlation of all intensities. We explicitly use Spearman correlation instead of Pearson due to its rank-based definition. Second, we analyze the industry rankings. For that, we compute the median intensity for each industry

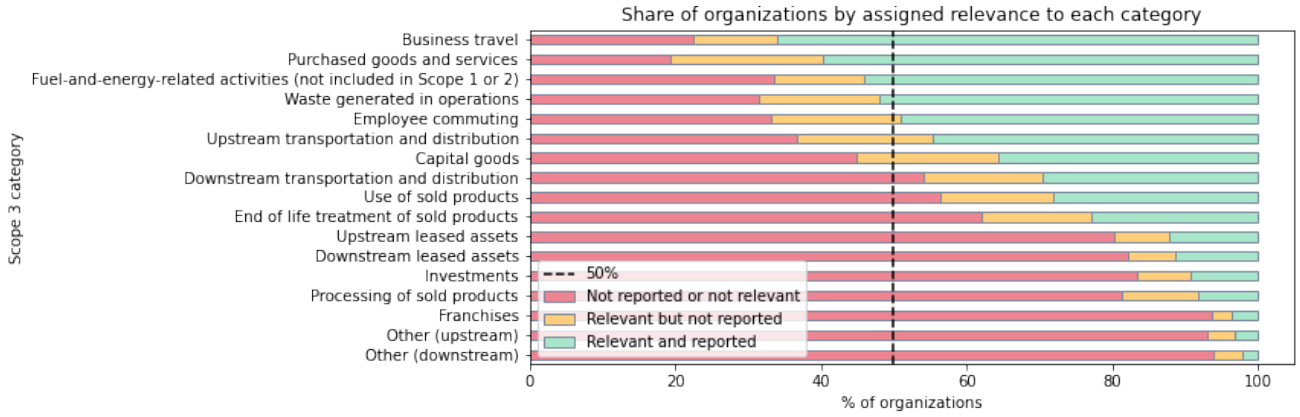

Figure 5. Share of organizations by assigned relevance to each category: we observe that most categories are not evaluated by most companies. Moreover, a significant share of companies recognize certain categories as relevant but do not evaluate them. Therefore, much information is likely still needed to be included in the CDP questionnaire.

and then report the Spearman correlation of the medians. Third, we measure to what extent the estimates and the reports are in similar orders of magnitude. We compute the share of industries for which the median estimated intensity is between the maximum and minimum values of the self-reported data points. Finally, to measure the ability to distinguish between firms in the same industry, we gauge the share of industries with a positive and significant within-industry Spearman correlation.

References

1. Arias, P. *et al.* Climate change 2021: The physical science basis. contribution of working group14 i to the sixth assessment report of the intergovernmental panel on climate change; technical summary (2021).
2. Tracker, C. A. Warming projections global update. november 2021 (2021). https://climateactiontracker.org/global/temperatures/, Last accessed on 2022-12-20.
3. Armstrong McKay, D. I. *et al.* Exceeding 1.5 c global warming could trigger multiple climate tipping points. *Science* **377**, eabn7950 (2022).
4. Matthews, H. S., Hendrickson, C. T. & Weber, C. L. The importance of carbon footprint estimation boundaries (2008).
5. Dai, R., Duan, R., Liang, H. & Ng, L. Outsourcing climate change. *European Corporate Governance Institute–Finance Working Paper* (2021).
6. Ilhan, E., Krueger, P., Sautner, Z. & Starks, L. T. Climate risk disclosure and institutional investors. *Swiss Finance Institute Research Paper* (2021).
7. et al, J. R. A corporate accounting and reporting standard. revised edition (2022). https://ghgprotocol.org/sites/default/files/standards/ghg-protocol-revised.pdf, Last accessed on 2022-12-20.
8. Martin Barrow, B. B., Carbon Trust. Technical guidance for calculating scope 3 emissions. version 1.0 (2022). https://ghgprotocol.org/sites/default/files/standards/Scope3_Calculation_Guidance_0.pdf, Last accessed on 2022-12-20.
9. Depoers, F., Jeanjean, T. & Jérôme, T. Voluntary disclosure of greenhouse gas emissions: Contrasting the carbon disclosure project and corporate reports. *Journal of Business Ethics* **134**, 445–461 (2016).
10. Andrew, J. & Cortese, C. L. Carbon disclosures: Comparability, the carbon disclosure project and the greenhouse gas protocol. *Australasian Accounting, Business and Finance Journal* **5**, 5–18 (2011).
11. Serafeim, G. & Velez Caicedo, G. Machine learning models for prediction of scope 3 carbon emissions. URL https://papers.ssrn.com/abstract=4149874.
12. Lyon, T. P. & Montgomery, A. W. The means and end of greenwash. *Organization & Environment* **28**, 223–249 (2015).
13. Wilson, M. C. A critical review of environmental sustainability reporting in the consumer goods industry: Greenwashing or good business. *J. Mgmt. & Sustainability* **3**, 1 (2013).

14. Tadros, H. & Magnan, M. How does environmental performance map into environmental disclosure? a look at underlying economic incentives and legitimacy aims. *Sustainability Accounting, Management and Policy Journal* (2019).

15. Haque, F. & Ntim, C. G. Environmental policy, sustainable development, governance mechanisms and environmental performance. *Business Strategy and the Environment* **27**, 415–435 (2018).

16. for Standardization, I. O. *Environmental management: life cycle assessment; Principles and Framework* (ISO, 2006).

17. Finkbeiner, M., Inaba, A., Tan, R., Christiansen, K. & Klüppel, H.-J. The new international standards for life cycle assessment: Iso 14040 and iso 14044. *The international journal of life cycle assessment* **11**, 80–85 (2006).

18. Guinée, J. B. & Lindeijer, E. *Handbook on life cycle assessment: operational guide to the ISO standards*, vol. 7 (Springer Science & Business Media, 2002).

19. Majeau-Bettez, G., Strømman, A. H. & Hertwich, E. G. Evaluation of process-and input–output-based life cycle inventory data with regard to truncation and aggregation issues. *Environmental science & technology* **45**, 10170–10177 (2011).

20. Lenzen, M. Errors in conventional and input-output—based life—cycle inventories. *Journal of Industrial Ecology* **4**, 127–148 (2000). URL https://onlinelibrary.wiley.com/doi/abs/10.1162/10881980052541981. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1162/10881980052541981.

21. Crawford, R. H. Validation of a hybrid life-cycle inventory analysis method. *Journal of Environmental Management* **88**, 496–506 (2008). URL https://www.sciencedirect.com/science/article/pii/S0301479707001272.

22. Downie, J. & Stubbs, W. Corporate carbon strategies and greenhouse gas emission assessments: the implications of scope 3 emission factor selection. *Business Strategy and the Environment* **21**, 412–422 (2012).

23. Crawford, R. H. Validation of a hybrid life-cycle inventory analysis method. *Journal of environmental management* **88**, 496–506 (2008).

24. Leontief, W. *Input-output economics* (Oxford University Press, 1986).

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

26. Suh, S. *Handbook of input-output economics in industrial ecology*, vol. 23 (Springer Science & Business Media, 2009).

27. Lenzen, M. Errors in conventional and input-output—based life—cycle inventories. *Journal of industrial ecology* **4**, 127–148 (2000).

28. Nakamura, S. & Nansai, K. Input–output and hybrid lca. In *Special types of life cycle assessment*, 219–291 (Springer, 2016).

29. Sodersten, C.-J. H., Wood, R. & Hertwich, E. G. Endogenizing capital in mrio models: the implications for consumption-based accounting. *Environmental science & technology* **52**, 13250–13259 (2018).

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

31. Yang, Y., Heijungs, R. & Brandão, M. Hybrid life cycle assessment (lca) does not necessarily yield more accurate results than process-based lca. *Journal of Cleaner Production* **150**, 237–242 (2017).

32. Patchell, J. Can the implications of the ghg protocol's scope 3 standard be realized? *Journal of Cleaner Production* **185**, 941–958 (2018).

33. Downie, J. & Stubbs, W. Evaluation of australian companies' scope 3 greenhouse gas emissions assessments. *Journal of Cleaner Production* **56**, 156–163 (2013).

34. Klaaßen, L. & Stoll, C. Harmonizing corporate carbon footprints. *Nature communications* **12**, 1–13 (2021).

35. Commission, E. Guidelines on reporting climate-related information (2019).

36. Friede, G., Busch, T. & Bassen, A. Esg and financial performance: aggregated evidence from more than 2000 empirical studies. *Journal of sustainable finance & investment* **5**, 210–233 (2015).

37. Hornuf, L. & Yüksel, G. The performance of socially responsible investments: A meta-analysis. *CESifo Working Paper* (2022).

38. Atz, U., Liu, Z., Bruno, C. & Van Holt, T. Does sustainability generate better financial performance. *A review, meta-analysis, and propositions. https://ssrn. com/abstract* **3708495** (2021).

39. Berg, F., Koelbel, J. F., Pavlova, A. & Rigobon, R. Esg confusion and stock returns: Tackling the problem of noise. Tech. Rep., National Bureau of Economic Research (2022).

40. Serafeim, G. & Velez Caicedo, G. Machine learning models for prediction of scope 3 carbon emissions. *Available at SSRN* (2022).

41. Davis, S. J. & Caldeira, K. Consumption-based accounting of co2 emissions. *Proceedings of the national academy of sciences* **107**, 5687–5692 (2010).

42. Dinda, S. Environmental kuznets curve hypothesis: a survey. *Ecological economics* **49**, 431–455 (2004).

43. Stern, D. I. The environmental kuznets curve after 25 years. *Journal of Bioeconomics* **19**, 7–28 (2017).

44. Babiker, M. H. Climate change policy, market structure, and carbon leakage. *Journal of international Economics* **65**, 421–445 (2005).

45. Aichele, R. & Felbermayr, G. Kyoto and carbon leakage: An empirical analysis of the carbon content of bilateral trade. *Review of Economics and Statistics* **97**, 104–115 (2015).

46. Essandoh, O. K., Islam, M. & Kakinaka, M. Linking international trade and foreign direct investment to co2 emissions: any differences between developed and developing countries? *Science of the Total Environment* **712**, 136437 (2020).

47. Essandoh, O. K., Islam, M. & Kakinaka, M. The precautionary principle. *World Commission on the Ethics of Scientific Knowledge and Technology (COMEST)* (2020).

48. Gallego, B. & Lenzen, M. A consistent input–output formulation of shared producer and consumer responsibility. *Economic Systems Research* **17**, 365–391 (2005).

49. Lenzen, M., Murray, J., Sack, F. & Wiedmann, T. Shared producer and consumer responsibility—theory and practice. *Ecological economics* **61**, 27–42 (2007).

50. Marques, A., Rodrigues, J., Lenzen, M. & Domingos, T. Income-based environmental responsibility. *Ecological Economics* **84**, 57–65 (2012).

51. Liang, S., Qu, S., Zhu, Z., Guan, D. & Xu, M. Income-based greenhouse gas emissions of nations. *Environmental science & technology* **51**, 346–355 (2017).

52. Ghosh, A. Input-output approach in an allocation system. *Economica* **25**, 58–64 (1958).

53. Leontief, W. *Input-Output Economics* (Oxford University Press, 1986). Google-Books-ID: HMnQCwAAQBAJ.

54. Nikaido, H. *Convex Structures and Economic Theory* (Elsevier, 2016). Google-Books-ID: NMVgDAAAQBAJ.

55. Hawkins, D. Some Conditions of Macroeconomic Stability. *Econometrica* **16**, 309–322 (1948). URL https://www.jstor.org/stable/1909272. Publisher: [Wiley, Econometric Society].

56. Kim, K. H. Introduction to sets and mappings in modern economics: Hukakane Nikaido, Amsterdam: North-Holland, 1970. *Mathematical Social Sciences* **2**, 331 (1982). URL https://www.sciencedirect.com/science/article/pii/0165489682910897.

57. Ghosh, A. Input-Output Approach in an Allocation System. *Economica* **25**, 58–64 (1958). URL https://www.jstor.org/stable/2550694. Publisher: [London School of Economics, Wiley, London School of Economics and Political Science, Suntory and Toyota International Centres for Economics and Related Disciplines].

Acknowledgements

The authors acknowledge David Cadrecha for his invaluable contributions to the development of the methodology.

Author contributions statement

Must include all authors, identified by initials, for example: A.A. conceived the experiment(s), A.A. and B.A. conducted the experiment(s), C.A. and D.A. analysed the results. All authors reviewed the manuscript.

Additional information

To include, in this order: **Accession codes** (where applicable); **Competing interests** (mandatory statement).

The corresponding author is responsible for submitting a competing interests statement on behalf of all authors of the paper. This statement must be included in the submitted article file.

Supplementary materials

Input Output summary

Wassily Leontief developed classical Input-Output economics in the 1930's[53]. This tool allows us to understand how goods and services flow through the different economic sectors [1] and countries until they reach the final consumer. In this framework, one can estimate the change in demand for industry inputs resulting from a change in production of the final good.

We compute the vector of economic outputs $\vec{x} = [x_i]$ of each sector $i \in \{1, \cdots, n\}$ in an iterative way. In the first step, the final output covers the final demand $\vec{f}$:

$$\vec{x}_0 = \vec{f} \tag{11}$$

To produce the vector of final demands $\vec{f} = [f_i]$, sectors themselves create demand. Therefore, each sector sells some of its output of goods and services to other sectors (intermediate output):

$$\vec{x}_1 = A \cdot \vec{f} \tag{12}$$

where $A = [a_{ij}]$ indicates the amount of output from industry i required to produce one dollar of output from industry j. In terms of dollars, A represents the industrial recipe for making one dollar of products and services belonging to each industry. Moreover, to cover industrial demand, sectors create a third level of demand:

$$\vec{x}_2 = A \cdot A \cdot \vec{f} \tag{13}$$

And thus adding up all the production levels, we can describe the final output in the economy as follows:

$$\vec{x} = (\mathbb{I} + A + AA + AAA + \cdots) \cdot \vec{f} \tag{14}$$

It can be shown that this geometric progression is convergent[54], and can be written as follows

$$\vec{x} = (\mathbb{I} - A)^{-1} \cdot \vec{f} = \mathscr{L} \cdot \vec{f} \tag{15}$$

where $\mathscr{L}$ is called the Leontief matrix [2]. The $\mathscr{L}_{ij}$ elements are interpreted as the total amount of dollars that pass through industry i for the final consumer to consume 1 dollar from industry j.

To compute A, we depart from the total sales volume or revenue (number of dollars) of intermediates going from supplier industry i to customer industry j, in the form matrix $Z = [z_{ij}]$, so that:

$$a_{ij} = z_{ij}/x_j \tag{16}$$

The full matrix of intermediates Z can be used then to state the following conservation identities:

$$x_i = \sum_j z_{ij} + f_i, \tag{17}$$

$$x_j = w_j + \sum_i z_{ij} \tag{18}$$

Equation 17 states that all output is used for production or consumption. Equation 18 on the other hand states that the difference between total output and consumption of inputs from suppliers is the gross value added vector $\vec{w} = [w_j]$. By dividing equations 17 and 18 by the output $\vec{x}$, we get two normalisation conditions:

$$1 = \gamma_i + \sum_j b_{ij}, \tag{19}$$

$$1 = \delta_j + \sum_i a_{ij} \tag{20}$$

where we have defined another matrix $B = [b_{ij}] = [z_{ij}/x_i]$, which contains the proportion of goods and services going from supplier i to customer j, and

$$\gamma_i = f_i/x_i, \tag{21}$$

$$\delta_j = w_j/x_j, \tag{22}$$

[1] A sector in this context can be any division of the economy: industries, countries, companies , business lines or a combination of the previous.

[2] Another way of deriving equation 15 is stating that total output is used in intermediate production plus final consumption, and thus $\vec{x} = A \cdot \vec{x} + \vec{f}$. Then, we can solve for $\vec{x}$ and arrive at the same expression.

Equations 17 and 19 represent that total output is consumed by final consumers or used by other producers. On the other hand, equations 18 and 20 state that total output is equal to the sum of inputs plus the value added. Some mathematical properties of this system have been proven in the literature. For any non-negative vector of final demand $\vec{f}$, there is a non-negative output vector $\vec{x}$ such that the economy is in balance. This is also known as the Hawkins condition[55] where the principal minors of the matrix $\mathbb{I} - A$ are all positive. The maximal eigenvalue of A satisfies $\lambda(A) < 1$, and the Leontief matrix exists and is non-negative[56].

Given the previous definitions, equation 15 describes economic output as a demand-driven system where the final demand determines the total output. However, we can also develop a supply-driven model[57]. In this case, total output is the sum of the final industrial value-added plus the value that one industry adds to the next, and so on. In a similar fashion than in Equation 2, we can describe the total production of the economy as:

$$\vec{x}^{\top} = \vec{w}^{\top} \cdot (\mathbb{I} + B + BB + BBB + \cdots) \tag{23}$$

Similarly as before, we can express equation 23 as:

$$\vec{x}^{\top} = \vec{w}^{\top} \cdot (1 - B)^{-1} = \vec{w}^{\top} \cdot \mathscr{G} \tag{24}$$

where $\mathscr{G}$ is called the Ghosh matrix. The $\mathscr{G}_{ij}$ elements are interpreted as the total amount of dollars that pass through industry i if industry j produces 1 dollar of value added.

Finally, and for future reference, we can compute the downstream process matrices as:

$$B = diag(1/\vec{x}) \cdot A \cdot diag(\vec{x}) \tag{25}$$

$$\mathscr{G} = diag(1/\vec{x}) \cdot \mathscr{L} \cdot diag(\vec{x}) \tag{26}$$

Where $diag(\vec{x}) = \mathbb{I} \odot \vec{x}$ and $\odot$ stands for the Hadamard or element-wise product .