Carbon emission and asset prices: new evidence from machine learning

Feng LI Xingjian ZHENG*
Shanghai Advanced Institute of Finance (SAIF)

January 20, 2024

Abstract

We estimate a large data panel of carbon emissions by US firms with a machine learning algorithm known as XGBoost. We predict scope 1 carbon emissions of listed firms from 2002 to 2021. This data set has a broad coverage of 4111 firms per year as compared to 1675 firms provided by data vendors. Based on this data set, we examine firms' carbon risk pricing in the US equity market. The result shows that the carbon premium was insignificant before the Paris Agreement, and the premium turned significantly more negative after the Paris Agreement. This contrasting phenomenon implies a positive shift in investors' ESG-related preferences and is more pronounced with our estimated data sample, where we provide a flow-based mechanism to explain the change in carbon premium. Overall, this paper complements rather than challenges previous empirical research from both sides and provides researchers with a novel approach to understanding climate finance.

JEL classification: G12, G23, G30.

Keywords: Carbon emission, Asset pricing, XGBoost, Paris Agreement.

*Li (fli@saif.sjtu.edu.cn) is from the Shanghai Advanced Institute of Finance at Shanghai Jiao Tong University and CAFR. Zheng (xjzheng.20@saif.sjtu.edu.cn) is from the Shanghai Advanced Institute of Finance at Shanghai Jiao Tong University. We thank colleagues from Finvolution Group Ltd. for their invaluable technical support. We benefited from extensive discussions and suggestions from Sona Agrawal (Discussant), Hui-Ching Chuang(Discussant), Claire Hong, Po-Hsuan Hsu, Qiushi Huang, Adrien-Paul Lambillon, Xiaomeng Lu, Shumiao Ouyang, Yuezhi Wu, and Chao Zi. All errors remain ours. Click *here* for the latest manuscript.

1. Introduction

There has been a longstanding debate on whether and how carbon risk is priced in the cross-section of expected returns. An influential research paper by Bolton and Kacperczyk (2021a) finds that stocks of firms with higher carbon emissions earn higher returns on average. However, others argue the emission-return relationship should be negative (Aswani et al., 2022; Garvey et al., 2018; In et al., 2017; Matsumura et al., 2014), and some even reach an inconclusive result (Monasterolo and De Angelis, 2020). One reason behind this controversy may be attributed to limited emission disclosure. Only a few firms voluntarily disclose carbon emissions, and researchers use carbon emission data from different data vendors for empirical estimation.

Currently, several data vendors provide researchers with carbon emission data (Busch et al., 2022), among which Bloomberg, Carbon Disclosure Project (CDP, and hereafter), ISS Ethix, MSCI, Sustainalytics, Thomson Reuters, and Trucost are the most widely used. These databases cover from scope 1 to 3 carbon emissions, or Greenhouse Gas emissions (GHG, and hereafter), for firms located in the US, EU, and other parts of the world from 2002 to the present. Notably, most firms included in these databases are located in the EU instead of the US, and US firms roughly take less than 30% of all the firms. Each database, on average, reports less than 2000 companies globally and less than 1000 companies per year for listed US firms. Many databases, led by Trucost Environmental, expanded their coverage after the Paris Agreement which was ratified on December 12^{th}, 2015, and the number of firms included in the database nearly tripled after 2016.

Besides, existing databases suffer from serious estimation biases. The estimated data provided by different data vendors, though it has a correlation coefficient ranging from 0.87 to 0.99 for scope 1 carbon emission (Busch et al., 2022), are primarily estimated by vendors instead of disclosed by firms themselves. Roughly 70% of the carbon emission data are from third-party estimations or simple forward-looking data. Their estimation algorithms seem to be a nearly deterministic linear function of size, sales growth rate, industry-fixed effects, and time-fixed effects (Aswani et al., 2022).

We argue that relying on an unbalanced data panel provided by data vendors would result in serious empirical problems, as there exhibits a non-negligible self-selection problem in disclosing carbon emission data, where only the "Clean/Green firms" with a high ESG awareness (and often good fundamentals) are willing to disclose carbon emissions voluntarily, which also on average are more profitable (Bolton and Kacperczyk, 2021b; Gibson et al., 2020; Görgen et al., 2020). Furthermore, the prediction algorithm applied by data vendors might be controversial, or even misleading to some extent, as most vendors opt for a linear function for emission estimation, whereas in reality, the real relation between carbon emission and other firm fundamentals or fixed characteristics like

industry and location appear to be highly non-linear.

In this paper, we seek to address this problem by predicting the carbon emissions of listed firms in the US equity market with a novel approach. We posit that firms that share similar business structures produce carbon emissions on a similar scale. We could infer the carbon emissions of a non-disclosure firm from its similar peers that have disclosed carbon emissions based on the business similarity score and other firm fundamentals[1]. We predict a large data panel from 2002 to 2021 of US-listed stocks with this methodology and find convincing evidence supporting the existence of a significant carbon premium. We document there was a slightly negative carbon premium before the Paris Agreement, and the premium drastically became much more negative afterward.

Following Hoberg and Phillp's pioneering research on cross-sectional business similarity (Hoberg and Phillips, 2010, 2016) and time-series similarity Cohen et al. (2020b), we identify how similar the two firms are by the Cosine Similarity Score, which originates from the Natural Language Processing literature. This widely used measure is computed with corpus from the business section of firms' 10-K reports, and it can quantitatively capture the similarity of firms' business structures. We benefit from the extensive coverage of similarity pairs in the US stock market from the database provided by Hoberg and Phillips (2010, 2016). The similarity score serves as an important independent variable to predict carbon emissions.

In addition to firm similarity, we include two unique features that may help predict firms' carbon emissions. The first measure is ESG awareness, which is extracted from firms' 10-k reports. We develop a method following Luccioni and Palacios (2019) and count the number of ESG-related words in 10-k to measure just how much the manager cherishes sustainable operations. The second unique feature is the number of green patents. We augment the WIPO green patent classification by analyzing whether this patent helps promote a more sustainable environment using the pre-trained large language model developed by Webersinke et al. (2021). This method helps to identify the green innovations (Leippold and Yu, 2023) and resolves the previously documented weak relationship between carbon emission and green innovation (Bolton et al., 2022c; Hege et al., 2023). These two unique variables are also key independent variables used to predict carbon emissions.

Next, we feed the cosine similarity scores, firms' carbon awareness, the number of green patents, and scope 1 carbon emissions of disclosure firms, and other firm fundamentals to a machine learning algorithm known as *XGBoost* to train the model. To address concerns related to the information leakage problem, we split the training and test sets by year instead of directly from the pooled samples. This partitioning method avoids the

[1]In the appendices, we use only firms' similarity scores as a predictor to show that this is indeed a valid argument. Moreover, when we assign random values to the similarity score pairs, the goodness-of-fit of the machine learning results declined drastically, implying that business similarity does help predict carbon emissions in our case.

overfitting issue with the traditional pooled sampling method, but it may not capture the time-varying component of carbon emissions. We select observations from 2002 to 2018 as the training set and observations from 2019 to 2021 in the test set. All carbon emissions are known for both sets when training the algorithm. After 2000 times of iterations, the model yields convincing results for both in and out-of-sample data.

Then, we apply the model to predict the carbon emissions of non-disclosure firms from 2002 to 2021. In the prediction set, for each similarity pair, only one of the two firms has carbon emission data, and the other firm is the non-disclosure firm. We predict carbon emissions and concatenate the data set with disclosed carbon emissions to merge into a full data panel. Since the disclosure of carbon emission in the year 2021 has not been fully disclosed by the Trucost database by the time we write this paper, we also do linear interpolation in the year 2021 based on emission data predicted with the XGBoost algorithm to enlarge the size of our database. Following this approach, we build a large data set, which consists of an average of 4111 listed firms per year, with the minimum firm-year observations of 2952 in the year 2002, and maximum observations of 4453 in the year 2021.

This panel is superior to the existing data set estimated by data vendors in two ways. First, our data set includes more firms before 2016, whereas most data vendors only include more listed firms after 2016. Second, we use a non-linear machine learning algorithm to predict carbon emission, which captures the non-linear relationship between firm fundamentals, industry-fixed effects, and time-fixed effects. Our method is not a static interpolation of disclosed carbon emission[2], as it includes the firm's business characteristics in the calculation.

We show that the data set we estimated is robust, and we design several empirical tests for more precise data validation. We first employ nationwide regulation shocks to examine firms' carbon emissions after a state announces an executive or statutory emission target. We expect a firm to experience a decrease in its carbon emission after a regulation shock. Regression results suggest that after a state announces its carbon emission target or has resolved to cut emissions, the firm would cut 61.62% of its carbon emission as compared to the control groups.

We also use a transition matrix to examine the persistence of carbon emissions. Intuitively, the carbon emissions of firms are highly serially correlated, as the tangible assets do not transfer or depreciate drastically over time. We first sort firms into five quintile groups based on their year 0 carbon emission intensity, and we report the probability that the firm should stay in this quintile group after 1/3/5/7 years. Empirical results suggest that the transition probability is quite stable, as roughly more than 70-80% of the firms

[2]Note that we only perform the linear extrapolation in the year 2021, as the disclosure dataset is incomplete for the moment. We are expected to update our prediction soon as the Trucost database fully updates the emissions in 2021.

stay in the same emission quintile after 1 year, and more than 60% of firms stay in the same quintile after 3 to 5 years. A similar analysis based on auto-correlation regressions following Bolton and Kacperczyk (2021a) in the appendices also supports the persistence of carbon emissions.

Besides, we examine the relationship between ESG fund inclusion and carbon emissions. We focus on ESG-related funds with investment objectives focusing on ESG factors. We identify ESG-related funds by searching for keywords such as "CLEAN", "ESG", or "SOCIAL" in their fund names. We regress the probability of a firm's probability of being included in an ESG-related fund on the logarithmic value of its carbon emissions, along with other firm fundamentals. Regression results suggest that the higher the carbon emission of a firm, the lower the probability its stock will be included in the portfolio of an ESG-related fund.

We also compare the determinants of firms' financial characteristics on their carbon emission in XGBoost-predicted data with the data provided by the Trucost database. Overall, regression coefficients are largely the same, which implies that the estimated sample by our machine learning algorithm captures the emission pattern that can be explained by firm fundamentals. However, a few variables like profitability ratio ROE and industry concentration ratio like the HHI index yield different contrasting results, which may be attributed to a large number of small-cap firms being included in the database before the Paris Agreement. Apart from empirical designs, machine learning validation results also support the robustness and credibility of the XGBoost-based results.

After validating the data set that we have estimated, we examine the pricing of firms' carbon risk in the US equity market. We conduct a battery of tests to examine the relationship between firms' carbon emissions and stock returns and to what extent is firms' carbon risk priced in the cross-section of stock returns.

Following Bolton and Kacperczyk (2021a), we regress monthly stock returns on firms' carbon emissions with three different measures. We test the emission-return relationship with either the data sample provided by the Trucost database or the sample estimated by the machine learning algorithms. In the first sample with the Trucost sample which spans from 2002 to 2016, we find there exhibits a significantly positive carbon premium, and the effect is more pronounced once we control for industry fixed effect. When we substitute the data sample estimated by XGBoost, the high-carbon premium becomes rather insignificant. The carbon premium is -0.0077 and -0.0090, with t-statistics of -0.83 and -0.92, suggesting the carbon risk is slightly negatively priced in the stock prices. However, when we further switch the sample period to that after the Paris Agreement (from 2016 to 2021), the premium significantly turns negative. The regression coefficients are -0.0798 and -0.0606 without and with the industry fixed effects, with t-statistics -2.39 and -1.81, respectively. The low-carbon effect is more pronounced (and consistent) when we use emission intensity, which is defined as carbon emission scaled by firms' sales, as

the independent variable of interest. These results suggest that investors have diverted their position towards less-carbon-intensive industries over the past few years, and the previously documented carbon premium might not be robust if we augment the Trucost dataset. In the appendices, we use linear models to predict firms' carbon emissions and re-examined the emission-return relationship. As shown in figure B1, the carbon premium estimated with linear models is roughly similar to the results estimated with Trucost data. This is because the linear models simply exacerbate the noise contained in firms' carbon emissions. When training the model, we show that the out-of-sample R2 prediction results by XGBoost significantly outperformed the linear models in figure A5. In other words, the XGBoost-based estimation not only outperforms the original dataset provided by the Trucost database but also linear estimation.

We provide reduced-form evidence by showing that the changes in carbon premium are at least partially driven by investor flow. We follow van der Beck (2021) by calculating the investor flow for stocks of firms with different levels of carbon emissions and show that investors are purchasing more green stocks after the ratification of the Paris Agreement, which pushed up the realized returns of carbon emission. In other words, the result we have documented is mostly "demand-based".

We further test the time-varying carbon risk premium and the impact of the well-known Paris Agreement with additional tests. Regression results in the appendices reveal that investors' preference for low-carbon stocks seemed to emerge only after 2012, and it was strengthened by a wake-up call of the Paris Agreement at the end of 2015 and became most significant in 2020. Regressions with common risk factors show that the shift in preference is not solely driven by risk preferences like size, earnings, margin investments, or liquidity. This relationship between asset prices and carbon emission, if not causal, is negatively correlated at least. In light of recent global or regional natural disasters such as drought, hurricanes, and extreme heat waves, investors gradually realize the importance of sustainable investment by selling carbon-intensive stocks (Alekseev et al., 2022; Choi et al., 2020a,b).

Interestingly, we provide additional evidence in table B4 suggesting that this shift is not solely driven by investors' attention but also by improvements in firm fundamentals. We show that post the Paris Agreement, firms with high carbon emissions are more profitable and become more financially stable. As a result, we may not rule out the alternative hypothesis that the low-carbon premium we observed after 2016 can also be attributed to fundamentally related risks. We conjecture that, once firms start to adopt advanced ESG-improving technology, they face higher operating costs because of more stringent production standards. Meanwhile, some firms may need to borrow from banks to upgrade their manufacturing lines, which raises the leverage ratio and makes firms more financially constrained.

Admittedly, the strong negative carbon premium after the Paris Agreement is less

pronounced for emissions measured by the growth rate, as the emission growth rate predicted by the algorithm is too volatile. This can be attributed to the estimation methodology: we are predicting carbon emission on a cross-sectional level even though we include firm fixed effects in the XGBoost model. Our methodology and data set, which mainly relies on the Trucost database, also is not exempt from the critique raised by Aswani et al. (2022). A huge portion of emission data in our original training set is generated by Trucost instead of disclosed by the firms themselves, and the unscaled emissions are either too correlated with firm fundamentals or clustered within industries with high heterogeneity. And yet, we argue that this paper at least provides researchers with a potential method to examine the pricing of carbon emissions before 2016, and also a possible way to examine the validity of emissions disclosed by firms or other data vendors in the future. For example, if the emission disclosed by firms is significantly lower than that predicted by the algorithm, say, three standard deviations away from the baseline prediction, then we have reasons to raise suspicions against the credibility of the emission data.

Overall, our empirical results can be summarized in one simple sentence. The positive emission-return relationship may no longer exist with more data, and the carbon premium is especially negative after the Paris Agreement. Moreover, prior studies tend to overestimate (but not falsely) the magnitude of the positive carbon premium as firms that disclose emissions voluntarily or are estimated by the Trucost database are larger and often belong to carbon-intensive firm groups, which may lead to biased estimation. Given that the result is largely driven by a positive shock for the ESG-related preference, we expect to observe the negative carbon premium reversed and become positively significant in the next few years.

It should be noted that we are not challenging the positive emission-return relationship documented in previous research especially led by Bolton and Kacperczyk (2021a,b). We successfully replicated their results in both the main results and the appendices with the Trucost dataset. Their analyses just may be limited to a lack of data and simply neglect the evolving interest in sustainable investment in recent years. We are mainly complementing their insightful results by reconciling with the views hosted by Pástor et al. (2021); Pedersen et al. (2021).

This paper contributes to the literature in the following ways. Firstly, this paper is the first paper that examines the carbon risk with machine learning estimated datasets. Previous literature that uses the Trucost dataset have contrasting views on the carbon premium (Aswani et al., 2022; Bolton and Kacperczyk, 2021a; Garvey et al., 2018; In et al., 2017; Matsumura et al., 2014; Monasterolo and De Angelis, 2020), this may be because of the lack of enough data. We complement their research by providing a larger dataset that has broad coverage of listed firms in the US markets and finds evidence supporting the findings in Zhang (2023), and this negative premium is mostly pronounced

after the Paris Agreement, suggesting a gradual increase in investors' carbon awareness (Acharya et al., 2022; Li and Zheng, 2024; Skiadopoulos et al., 2023). Moreover, this paper is unique in the sense that we document a negative premium with the "level" of carbon emission whereas the result is insignificant with previous research.

Secondly, this paper contributes to the literature that proposes a reliable method to predict carbon emissions. Previous research relies on more rudimentary machine learning algorithms like gradient boosting decision trees (Han et al., 2021). Our extreme GBDT method outperforms their method due to XGBoost's advantages in regularized feature splitting and more efficient tree pruning. Other work adopts RBF, Elastic Net, or meta-analyses to predict carbon emission (Javadi et al., 2021; Mardani et al., 2020; Nguyen et al., 2021). Our prediction is different from theirs in the sense that our method is built on robust economic premises and therefore, more suitable for asset pricing research. Besides carbon emissions, researchers could also use this approach to predict ESG scores, patents, and other important variables that have strong industry-fixed effects. It may also be useful to check whether there might be a false statement of carbon emissions disclosed by firms in the future. Unfortunately, this method based on XGBoost is only efficient in predicting scope 1 emission data, as emissions of other metrics cannot be only captured by business structures, especially for scope 3 emissions.

The rest of the paper is organized as follows. Section 2 discusses related literature from three perspectives, including economic links between firms, using regression trees in Econ and Finance, and the implications of carbon disclosure. Section 3 describes our method to predict carbon emissions with a battery of robustness tests. Section 4 examines the carbon premium and the pricing of common risk factors. Section 5 concludes.

2. Related literature

2.1. Economic links and industrial competitions

There has been extensive research on identifying similar or related firms with textual analysis in recent years. Hoberg and Phillps pioneer the work by analyzing firm 10-K product descriptions in the Business sections Hoberg and Phillips (2010, 2016, 2018). They create word vectors and compute similarity scores between two firms, and their method generates a new set of industries in which firms can have their own distinct and time-varying set of competitors. Their method is superior because it allows researchers to examine how close two firms are in a vector space with a continuous variable instead of common industry categories like SIC codes or NAICS codes. Cohen et al. (2020b) followed their approach and used other similarity measures to compute time-series similarity for a firm itself.

Apart from creating word vector space, there are simpler ways to identify industry

competitors or allies within industries. Li et al. (2013) counts the number of times a firm refers to competition in its regulatory 10-K to measure a firm's competing environment, which behaves as if it is measuring the "true" competition. This measure is also adopted by Bustamante and Frésard (2021), Bernard et al. (2020), and Eisdorfer et al. (2022) for its simplicity.

Other researchers use different and intriguing methods to measure firm links. Lee et al. (2019) examined the technological linkage between the two firms by exploiting various categories in the patent data and calculating a pairwise measure of technological closeness. Cohen and Frazzini (2008) extracted firms' customer information from segment files between 1980 and 2004. The economically related firms between suppliers and customers have strong predictability of future stock returns. Other firm-level links worth noting include the common analyst coverage link (Ali and Hirshleifer, 2020), social ties (Peng et al., 2022), CEO's personal connections (Engelberg et al., 2012, 2013), and geographical links (Jin and Li, 2020).

2.2. Boosting trees in Economics and finance

The *Extreme Gradient Boosting* algorithm is an advanced machine learning algorithm ensembled on gradient boosting, and it was developed by Chen and Guestrin (2016). XGBoost has an additive feature that is trained at each iteration, and it is highly efficient when the data set scale is on the order of 100 thousand to 1 million.

The Econ-and-finance literature mainly uses XGBoost or other boosting models for classifications, as it has superb performance for pushing the limits of computing power for boosted tree algorithms. With XGBoost, classifications or predictions could be built in parallel by splitting data samples in the training set. In Zheng (2022), he trains a machine learning algorithm using earlier patent applications and predicts the good-or-bad quality of recent applications out of sample. The training results are promising, resulting in a 15.5% gain of patent generality and a 35.6% gain in the number of patent citations. Another application is within the field of consumer finance. Tantri (2021) used XGBoost to improve the efficiency in lending without leading to an increase in default in an Indian Bank. The result suggests that, with the help of the algorithm, lenders can financially include 60% more at loan officers' delinquency rate or achieve a 33% lower delinquency rate. Other studies applied this model directly to asset returns. Teng et al. (2020) applied several machine learning algorithms from random forests to neural networks. They find that when building a buy-and-hold portfolio, XGBoost, and Neural networks produce portfolios with the highest Sharpe ratios. Other researchers use traditional boosting algorithms to examine the cross-sectional variation in the effects of Robo-advising on retail investors' portfolio allocations and performance (Rossi and Utkus, 2020). Undoubtedly, researchers can also use other machine learning algorithms

such as Neural Networks, Random Forests, SVMs, or even simpler logistic regressions for emission prediction, but we only discuss methods with the boosting trees for simplicity and efficiency.

2.3. Carbon emission and stock returns

Finally, we contribute to the literature on carbon disclosure, the financial cost of carbon disclosure, and the cross-section of stock returns. As we have discussed, there has been a controversial debate on whether and how carbon emission is priced in stock returns. Some researchers document a positive link between stock returns and emissions, as led by Bolton and Kacperczyk Bolton et al. (2022a); Bolton and Kacperczyk (2021a,b); Bolton et al. (2022b); Bolton and Kacperczyk (2020a,b, 2021c); Bolton et al. (2022c, 2021), and Ilhan et al. (2021), which is consistent with the risk compensation hypothesis. The idea behind this hypothesis is quite straightforward, as firms with higher greenhouse gas emissions are more vulnerable to state environmental penalties or other environmentally related risks. As a result, investors require a higher rate of return for extra risk compensation. A similar study targeting the pollution premium has similar results, in which investors demand pollution-related risk compensation. They prove that firms with higher pollution are more sensitive to litigation risk (Hsu et al., 2022).

However, some other researchers believe this phenomenon is entirely driven by vendor-estimated emissions, which makes the estimation results quite unreliable (Aswani et al., 2022). Duan et al. (2021) examine the pricing of a firm's carbon risk in the corporate bond market and find that bonds of more carbon-intensive firms earn significantly lower returns than their industry peers. Their empirical results are more robust because multiple existing bonds exist for a single firm, making time-series estimation available. In Cheema-Fox et al. (2021), researchers construct a decarbonization factor that goes long low-carbon intensity firms and shorts high-carbon intensity firms. This decarbonization factor yields significantly positive returns, especially in Europe. In a recent analysis with global evidence Choi et al. (2022), researchers find that high-emission firms tend to have lower price valuation ratios than low-emission firms, and the devaluation of high-emission firms phenomena are most prominent in recent years. Their empirical analyses mainly focus on the valuation ratios such as PE, PS, and PB, instead of stock returns.

This low-carbon premium might also be attributed to raising awareness of ESG investing (Pástor et al., 2021, 2022; Pedersen et al., 2021) in recent years. This is consistent with increasing evidence documenting that institutional investors around the globe have started to decrease their portfolio exposure towards high carbon emission firms (Bolton and Kacperczyk, 2021a; Choi et al., 2020b, 2022; Gibson et al., 2020). In van der Beck (2021), he shows that the cumulative total flows into the ESG portfolio increased 7 folds, from 0.2 trillion US dollars in 2017 to 1.4 trillion in 2022. A huge amount of money

has been diverted from the market portfolio towards the ESG portfolio, pushing up the realized returns (but not the expected returns). Another explanation could be that lower or reduced carbon emission ratios are associated with stronger future profitability and positive stock returns in a global universe of stocks (Garvey et al., 2018; Görgen et al., 2020). Mutual fund managers who specialize in responsible investments exploit this underreaction of mispricing and high profitability (Glossner, 2021).

Overall, the literature from both sides, which adopt different empirical methodologies and data samples, gives inconclusive results and demands more empirical validation, and nearly all prior research is limited to a lack of emission data.

3. Greenhouse Gas data estimation

3.1. Data and variables

We estimate carbon emissions on the basic premise that firms that share similar business structures produce carbon emissions on a similar scale. We can use the business similarity and carbon emissions of firms that have disclosed carbon emissions to predict the carbon emissions of undisclosed firms.

We present a simple example in figure 1 to illustrate this premise. Suppose in the automobile manufacturing industry, three firms, Ford, Toyota, and Tesla, have disclosed their carbon emissions, whereas General Motors does not. In the first stage, we fit an emission-business similarity relationship within the disclosed group. We can thus know how business similarity within the automobile industry is related to emission similarity. Then, in the second stage, we use the fitted relationship and the disclosed carbon emissions of Ford, Toyota, and Tesla to predict the undisclosed carbon emission of GM, since GM is closely related to the other three firms.

[Insert Figure 1 near here]

Following this simple conjecture, we input the firm business similarity score, scope 1 GHG (Greenhouse Gas emissions), and other firm fundamentals into the XGBoost algorithm. The firm similarity pair score is obtained from the Hoberg and Phillips original Data Library (Hoberg and Phillips, 2010, 2016). This data set can be traced back as far as 1989 and is updated until 2021 on a biannual basis, covering the majority of listed firms in the US stock market. We use the baseline TNIC similarity data. We selected 2002 as the start of our test period because our carbon emission data started in 2002, and the latest TNIC similarity data ended in 2021 by the time we downloaded data from Hoberg and Phillps's data library. The similarity score is highly skewed to the right, and 75% of the scores are lower than 0.2. As a result, for each firm, we sort similarity scores from most similar to least similar and keep its top 20 similar firms so that we can filter

noise, as many firm pairs have similarity scores close to zero. For example, the business similarity between GM and a pharmaceutical company is nearly zero, and adding this pair of observations into the algorithm hardly helps train the model. We present summary statistics of the similarity scores by year in table 1. We can also expand the selection threshold from 20 to 30 or more to include more firms in the algorithm for more prediction, but the marginal contribution is very limited.

[Insert Table 1 near here]

We obtain carbon emission data from the Trucost database, which is widely adopted by previous research, most notably led by a series of insightful works by Bolton and Kacperczyk. This database provides researchers with three scopes of carbon emissions from scope 1 to scope 3. According to the definition, Scope 1 emissions are direct emissions from company-owned and controlled resources. All fuels like gas, oil, and electricity that produce GHG emissions must be included in scope 1 emission. Scope 2 emissions are indirect emissions from the generation of purchased energy from the firm's utility provider and their consumption of purchased electricity, steam, and heat. Scope 3 emissions are indirect emissions not included in Scope 2 that occur in the reporting company's upper or lower value chain. We mainly use scope 1 GHG emission in this paper, as it directly measures the real emission produced by the PPEs of a firm. We ignore scope 2 or 3 GHG emissions because they are either indirect emissions or more suitable for financial firms. Using scope 1 GHG to measure carbon emissions produced by industrial manufacturers or other non-financial companies is more appropriate. Importantly, data vendors tend to make frequent updates, and more firms are included in the database than previously documented. In 2023, they are using an estimation methodology where emission reports are not available for previous years and thus expanding the database on a larger basis.

We include firm characteristics, including size, total assets, non-current assets, and employee numbers, and all variables are obtained from the Compustat database. We winsorize all firm fundamental variables at 2.5% on both tails to remove outliers.

Additionally, we include two unique features that measure firms' idiosyncratic environmental attributes. The reason is that different firms often exert different degrees of effort to promote a sustainable business, and these efforts are empirically difficult to identify with only firm fundamentals and industry-fixed effects. To capture the idiosyncratic green component, we add two measures into the XGBoost algorithm. The first component is firms' ESG awareness, and the second component is the number of firms' green patents.

The first variable, ESG awareness measures firms' environmental consciousness. Firms with higher ESG awareness are more likely to employ advanced emission reduction technologies, resulting in lower carbon dioxide emissions and better ESG performance. If the managers are more aware of the imminence of climate change, they will disclose more en-

vironmental or ESG-related information in their annual reports(Matsumura et al., 2022). In that sense, we search for ESG-related keywords and count the number of keywords in firms' 10-k to proxy for firms' ESG awareness. To do so, we follow Chang et al. (2023); Luccioni and Palacios (2019) by manually specifying a set of seed words related to the three aspects of ESG, feeding these seed words to the Word2vec algorithm to retrieve the thirty closest words in the text using cosine similarity and obtain a full-fledged ESG dictionary. Finally, we span a dictionary that has 182 ESG-related words in total.

The second variable is the number of green patents. Previous literature documents that scope 1 and scope 2 carbon emissions may not be related to firms' green patents (Bolton et al., 2022c), but are related to firms' scope 3 emissions (Hege et al., 2023). However, other research documents a significant association between the number of firms' green patents and asset prices and firm profitability (Leippold and Yu, 2023) and the association between firms' ESG-related innovation between carbon emissions may be different quite surprising (Cohen et al., 2020a). One reason behind the insignificant result may be the coarse classification of green patents. We resolve this problem by augmenting the original WIPO green patent classification with advanced NLP methods developed by Webersinke et al. (2021). We input the patent abstract collected from Stoffman et al. (2022) into a pre-trained language model known as Climatebert. We combine the predicted patents by Cliamtebert and the original classified green patents by WIPO together as green patents and define a new variable $GREENPATENT$ that measures the number of green patents a listed firm has been granted in a year.

3.2. GHG prediction with XGBoost

We use a regression tree method, XGBoost, to predict greenhouse gas emissions GHG. This algorithm does not only depend on conditional linear estimation but also other non-linearity features. Besides, XGBoost is a decision tree ensemble based on tree boosting, one of the most popular supervised machine learning algorithms in industry and academia. We choose this algorithm due to its superior performance over other scalable machine-learning models in solving regression, classification, and ranking problems (Basu et al., 2023; Tantri, 2021; Zheng, 2022).

The independent variable of the regression tree (prediction) is the scope 1 carbon emission or the Green House Gas (GHG) emission of the non-disclosure firm defined as GHG_f. The independent variables of the regression tree include the GHG emission of the disclosure firm GHG_d, the cosine similarity between the undisclosed firm and the disclosed firm, and fundamentals including sales, total assets, employee numbers, and non-current assets of both firms. Since the carbon emission of firms is serially correlated, we include the GVKEY of the non-disclosure firm into the model and set this variable as a categorical value. This method could help us predict the carbon emissions of the

non-disclosure firm more accurately and consistently. Finally, the model trained on a cross-sectional level can be expressed as follows.

$$\widehat{GHG}_f = \hat{f}\left(GHG_d, \text{score}_{<f,d>}, \cdots\right) = \arg\min L\left(f(X) + \widehat{GHG}_f\right) + R(f(\cdot)) \quad (1)$$

We use a five-fold cross-validation test to prevent over-fitting issues and train the model at 2000 times iterations. We have 237640 firm similarity pairs with GHG and other firm fundamentals. We manually split this data set based on the observation year. We select observations from 2002 to 2018 as the training set and from 2019 to 2021 as the test set, where there are 151348 observations in the training set and 86292 observations in the test set, respectively. We use the XGBoost algorithm trained by the training set to predict the carbon emissions of firms that did not disclose carbon emissions. For the prediction set, we have 363574 observations, where we can use disclosed emissions, firm similarity scores as well and other variables to predict the carbon emission of the undisclosed firms. We do not introduce all the machine learning parameters in this paper for brevity, but we report robustness and validation tests in the appendices.

[Insert Figure 2 near here]

We present an illustrative example of model training in figure 2, where we partition data into a training set, a validation set, and a prediction set. The dependent variable is on the left-hand side of the figure, where firms do not disclose their carbon emissions. On the right-hand side of the figure are the independent variables, including similarity scores, GHG from the disclosure firms, other firm fundamentals like the logarithmic value of firm sales, total assets, non-current assets, and employees for both the non-disclosure firm f and the disclosure firm d, and a firm-fixed dummy of the non-disclosure firm. We report summary statistics of the training and validation set in table 2. This table reports pooled observations of both the training set and the test set. The pooled sample period is from 2002 to 2021. We take the logarithmic value of all the firm fundamentals for emission prediction, whereas we use real values for prediction when training the algorithm. The standard deviation of the GHG is large because we are using raw scope 1 carbon emissions for summary statistics.

[Insert Table 2 near here]

The estimated model yields convincing results, where after 2000 times of iterations, the learning curve for the train set and the test sample set remain stable. In figure 3, subfigure A, we plot the learning curve with the valuation metric Root-mean-square-deviation (or RMSE) for both in-and-out-of-sample model training. The Root Mean Square Error shrinks drastically after 200 times of iteration for out-of-sample training,

and the curve becomes flatter after 1000 times. In the appendices, we also try different sample periods, where we partition sample data by splitting the sample with a training period from 2002 to 2016 and a test set from 2017 to 2021, or the training set from 2002 to 2019 and a test from 2020 to 2021, respectively, and re-run the XGBoost algorithm. We report detailed training results, and different partitioning periods yield similar results.

In subfigure B of figure 3, we plot the relationship between estimated carbon emissions versus real emissions for out-of-sample data, i.e., the test set. The x-axis is the predicted emission for the test set, and the y-axis is the real emission for the test set. We add a 45-degree line for better visual illustration. As can be seen from the figure, the out-of-sample model produces a remarkable fit and the R-square is 0.81, with most of the dots concentrated around the 45-degree line, which suggests that our estimated data is a good fit for real values. There are a few horizontal outliers for high-emission firms, and the fitted line is not perfectly located on the diagonal position in the plot, suggesting that our model might potentially underestimate real carbon emissions for brown firms. We conduct a battery of robustness analysis in the next section.

[Insert Figure 3 near here]

We also perform an importance plot (or the relative influence plot) following previous research (Medina and Pagel, 2021; Rossi and Utkus, 2020) in subfigure A of figure 4. This figure illustrates the importance of each variable in predicting carbon emissions. The result is quite intuitive, where the first two influential variables are the cosine similarity score and the carbon emissions of the disclosed firm. The next few important variables of interest are the non-current asset, sale, a firm-level fixed dummy, and employee numbers for the non-disclosure firm. Firm fundamentals of the disclosed firms contribute the least to this model. It is important to note that the unique dummy variable that identifies the firm proves to be useful, as the pattern of carbon emission is quite consistent over time. Moreover, ESG awareness and ESG patents are among the last important variables, which is because most of the values of these two unique features are zero and contribute relatively less as compared to the other variables. Overall, the importance plot suggests that our method can capture the serial correlation in carbon emission.

A more illustrative plot is the SHAP (SHapley Additive exPlanations) value plot, which is a game theoretic approach to explain the output of any machine learning model in subfigure B of figure 4. The higher the SHAP value, the more important the variable contributes to the model. This measure is more widely adopted than traditional importance plots in the finance literature (Erel et al., 2021). As can be seen from this figure, firm fundamentals of the non-disclosure firms increase model prediction performance, pushing the prediction away from the baseline value. The firm-level fixed variable GVKEY is also illustrated in this plot, and not surprisingly, its impact distribution is symmetric around the baseline vertical line. As shown in the appendices (see A1), the GVKEY proves to be

probably the most important predictor of firms' carbon emissions. In unreported results, we find that with this fixed identifier along, the prediction accuracy can achieve as high as more than 60%. This figure also shows that the firm sales, with a few outliers clustered to the right, help predict high greenhouse gas emissions. In other words, some high-sale firms produce extremely high carbon emissions. Additionally, the ESG awareness of the undisclosed firm also contributes heavily to emission prediction, which contrasts the results in the importance plot. After training the model and reporting basic validation results, we apply the algorithm to the prediction set, where only one side of the similarity pair has disclosed carbon emissions. We set a lower bound of prediction as zero to avoid negative predictions. Besides, since we perform the prediction on a cross-sectional level, we make linear interpolation in 2021 if the carbon emission is still unavailable after the prediction to enlarge and balance the dataset at full potential.

[Insert Figure 4 near here]

3.3. An overview of the carbon emission data

We report summary statistics of our estimated data set in table 3. Our average data contains 4111 firms per year, including financial firms with 2-digit GIC code 40 in three major US stock markets. We compare our data with data from other vendors, including the original Trucost data, the CDP (Carbon Disclosure Project), and Thomson Reuters. For the Trucost data obtained from the WRDS database, we match firms with GVKEYs. For the CDP and Thomson Reuters data, we match firms with their CUSIP IDs. We report the number of disclosed firms in their database annually in columns 1 to 3. It can be seen that our data outnumber data provided by other vendors, especially before 2016. We also report the empirical data used in Aswani et al. (2022), which replicates Bolton and Kacperczyk (2021a) with Trucost data. The sample period of their paper begins in 2005 and ends in 2019, and it has 1176 stocks on average per year.

[Insert Table 3 near here]

We compare the estimated data set with the original data set obtained from the Trucost database in table 4 in detail. For each data set, we report summary statistics of its logarithmic greenhouse gas emission by year from 2002 to 2021. Overall, the estimated dataset is comparable to Trucost's original emissions in magnitude after 2016, with an average logarithmic emission of 10.12 and a median logarithmic emission of 10.65, respectively. The standard deviation of our dataset is slightly larger than the original data, as our database contains many more firms. Before 2016, the Trucost database only included emissions of large firms or heavy pollution firms, which drastically increased the

average emission. The number of disclosed firms estimated by XGBoost in 2021 is 4453 because we have performed linear interpolation based on estimated data.

[Insert Table 4 near here]

We also report carbon emissions estimated by the algorithm by industry in table 5. We report 2-digit GIC industry classification emissions on the firm-year level. We sort industries based on average firm emissions. The highest emission industry is utilities, followed by materials and energy. These three industries tend to emit massive greenhouse gases when producing basic products or services. Industries that emit the lowest emissions are financial services, real estate, and information technologies, which intuitively do not involve heavy manufacturing. We report detailed carbon emissions and the number of firm-year observations in panel B with 6-digit GIC codes. Similar to results in panel A, utilities and transportation companies emit the most, whereas consumer finance, banks, and thrifts & mortgage finance companies emit the least. However, the GIC industry classification method is vulnerable to the coarse classification problem, which renders limited firms in certain categories. Moreover, we are reporting greenhouse gas emissions using the scope 1 metric. At the same time, financial firms like banks tend to have higher emissions on the scope 3 metrics which considers firms' upstream suppliers or downstream customers. Since our estimation method is mainly based on business similarity, it may not be reliable to apply the same methodology to scope 2 or scope 3 metric emissions. Scope 2 emissions are "indirect" emissions created by the production of the energy that the firm purchases. Business similarity may not imply similar energy consumption from power plants. On the other hand, scope 3 emissions are carbon emissions produced by customers using the firms' products or those produced by suppliers making products that the firm uses. This involves detailed similarity information on the firms' suppliers or customers within one industry and it requires more industry-specific knowledge.

[Insert Table 5 near here]

Following Bolton and Kacperczyk (2021a), we report summary statistics of firms' carbon emission, along with other firm fundamentals that will be used in empirical estimations, in table 6 panel A. Emission data include the logarithmic value of carbon emissions, carbon emission intensity defined as raw emission scaled by firm sales. Firm fundamentals include firm size at the end of the year, leverage ratio defined as the book value of debt divided by assets, investment ratio defined as CAPEX over book value of total assets, return on equity, HHI Herfindahl index of the industry, the logarithmic value of plant, property & equipment LOGPPE, book-to-market ratio defined by the book value of equity divided by market value of equity, sales growth and EPS growth normalized by last years value. In the next sub-panel, we report year-month level variables, including monthly stock returns, stock momentum defined as the cumulative stock

return over the last 12 months and excluding the last month, volatility as the standard deviation of return over the last 12 months, and the CAPM BETA calculated over the last 24 months. The summary statistics of panel A include min, 25%/50%/75%, and the max value of all the variables. We winsorize the carbon emission at 2.5% level at both tails to exclude extreme carbon estimation and other financial variables. The mean value of logarithmic carbon emissions is 11.07 with a standard deviation of 3.62, and the max value is 15.92. The distribution of carbon emissions generally is skewed to the right as few large manufacturing and petrol chemical firms produce emissions heavily. Moreover, similar to the level of carbon emissions, emission intensity is highly skewed as well. The emission intensity has a maximum value of 703.62, suggesting there exist small firms that produce significant amounts of carbon emissions. In panel B, we report the correlation matrix between carbon emission between other firm fundamentals. After taking logarithmic values, the value of correlation coefficients dropped significantly. Summary statistics suggest that the correlation matrix suggests that emission is not correlated with firm financing constraints or profitability ratios like Roe. Moreover, this variable is more correlated with firms' plant, property & equipment because this measure suggests how much this business operation relies on tangible assets.

[Insert Table 6 near here]

3.4. Robustness tests for the data set

In this section, we conduct a battery of validation tests to ensure our method predicts convincing results. We would like to examine the validity of our data set before performing any further asset pricing tests. Our first analysis relies on state-level regulatory changes regarding carbon emissions. In 2005, California pioneered reducing carbon emissions by then-Governor Arnold Schwarzenegger, and many states followed California's path. Until the end of 2022, 23 states plus the District of Columbia have announced emission targets to address climate change. These policies include carbon pricing, emission limits, renewable portfolio standards, and steps to promote cleaner transportation. Among all the policy targets, we manually collect a target mostly related to our topic and usually, the first to be announced by the state: carbon emission targets. Our policy data comes from C2ES or the Center for Climate and Energy Solutions.

Among the 23 states that have announced carbon emission targets by the end of 2021, 18 released announcements after 2017 (one year after the Paris Agreement). One thing to be noted is that California did not only pass one policy. We summarize their detailed state policies in the appendices.

Our identification strategy is very similar to a staggered DID, as the policy shock occurs in different states and in different years. We investigate firms' carbon emissions

in these "Green states" before and after the policy shock. Admittedly, our identification strategy is imperfect as there are strong state-fixed effects. Intuitively, states controlled by the Democratic parties usually announce more emission targets than the states controlled by the Republican parties. Moreover, the emission pattern within a state is very likely to be clustered at the industry level. Two notable examples or comparisons are the states of California and Texas, which are long to be conceived as deep blue and deep red. Texas has yet to announce a carbon emission target in 2022 because a large fraction of the firms within Texas depend on high-carbon petrol and chemical firms. Our regression formula is as follows.

$$LOGGHG_{i,t} = \alpha + \beta REGU_{i,t} + \gamma' X_{i,t} + \mu_j + \lambda_s + \varepsilon_{i,t} \tag{2}$$

The dependent variable $LOGGHG$ is the logarithmic value of the firm's scope 1 greenhouse gas emission. The independent variable $REGU_{i,t}$ includes a dummy variable that indicates whether the firm's state has experienced a carbon emission shock or a continuous variable that indicates years before or after the carbon emission shock. If the variable is negative, the state has yet to announce a carbon emission target or set out force to reduce carbon emission. Control variable $X_{i,t}$ includes the firm's size $LOGSIZE_{i,t}$, book-to-market ratio $B/M_{i,t}$, leverage ratio $LEVERAGE_{i,t}$, investment ratio $INVEST2A_{i,t}$, ROE ratio $ROE_{i,t}$, Herfindahl index $HHI_{i,t}$, natural logarithmic value of plant, property & equipment $PPE_{i,t}$, sales growth $SALESGR_{i,t}$, and EPS growth $EPSGR_{i,t}$. We include the firm's state fixed effect λ_s, its 6-digit GIC industry fixed effects μ_j, and $\varepsilon_{i,t}$ denotes residuals. All standard errors are clustered at the industry level. The regression variable of interest is the dummy variable, and the continuous variable indicates regulatory shocks. If the coefficient is negative, then it implies the firm cut emissions under the state's policy pressure. The regression results are illustrated below, where t-statistics are displayed below the coefficients. Regression results are shown in table 7.

[Insert Table 7 near here]

Table 7 shows the regression coefficients on the dummy and continuous variables are significantly negative (the first and the second row). As seen from the first column, after a state announces its carbon emission, the firm would cut 0.6162% of its carbon emission intensity, which is economically significant. Interestingly, the effect is more pronounced after we control for the state-fixed effect, implying that our XGBoost-based emission data somehow magically captures the policy tendencies across states. Adding control variables does not shrink or change the negative impact of regulatory shocks on firms' carbon emissions. In column 4, regression results suggest that a firm would cut down 0.1038% of its carbon emission 5 years after the state announced an emission target. This effect is considerably strong, as the firms' emissions tend to positively co-move with their

revenues and size. However, the negative relationship suggests that firms would adopt clean technology to tackle the regulatory shock. This result is in line with the recent study by Tomar (2022), where the disclosure of carbon emissions induces firms to reduce their emission (through benchmarking).

Next, we explore carbon emission persistence in our data set. Since firms do not adjust their business structures frequently, carbon emission relies heavily on plants and equipment and should not be very volatile. As a result, carbon emissions are quite persistent. We compute the carbon emission transition matrix for firms belonging to different quintiles. We follow the empirical method by Hsu et al. (2022) and sort firms into five quintiles based on their logarithmic carbon emission computed by dividing firms' greenhouse gas emission by firms' sales at year 0 and assigning them to five carbon quintiles 1/3/5/7 years after year 0. The results are displayed in table 8. The results are similar if we use other carbon emission intensity measures by dividing raw emission by other firm fundamentals like sales, non-current assets, total assets, and PPEs.

[Insert Table 8 near here]

As can be seen from table 8, the transition matrix is quite stable. For firms assigned to group 1 (the lowest carbon emission group) at year 0, the probability that it remained in group 1 is 80.68%. For firms assigned to group 5 (the highest carbon emission group) at year 0, the probability that it remained in group 1 Is 76.74%. Transition probabilities across groups shrink by year, which is reasonable because firms may adjust their operation units by turning to a carbon-intensive or carbon-reducing style. Also, new firms may enter the samples, which would crowd out firms from the original groups. In the appendices, we follow Bolton and Kacperczyk (2021a) by performing auto-correlation coefficients for three different Scope 1 greenhouse gas emissions. Regression results suggest that the logarithmic value of emission and emission intensity is quite persistent, whereas the emission growth rate is not. The non-persistent growth rate may be because we are predicting the growth rate on a cross-sectional level and performing linear interpolation to maximize the observation number. It would not bias our baseline estimation with logarithmic emission.

We also consider another robustness analysis by investigating mutual fund holdings. It is widely perceived that green funds or ESG funds can pick stocks with better ESG performance. They do so not only because their investment objectives mandate to do so but also because fund managers can filter green stocks by various measures. Some mutual fund managers inquire MSCI ESG index; others attend on-site roadshows. We identify ESG-related funds by searching for keywords including "ESG", "CLEAN", and "SOCIAL" in fund names. We define a fund as an ESG-related fund if it contains any of the three keywords in the fund name. Note that this is a smaller set of ESG dictionaries and we can definitely expand by including more words or use methodologies

such as Word-to-vec for expansion. We also exclude ETFs by excluding funds that hold more than 200 stocks in their position each quarter and run the following regression, where the dependent variable is either the total number of being included in ESG-related funds portfolio $TotalInclusion$ or its probability of being included in an ESG-related fund $InclusionProb$. The dependent variables include logarithmic carbon emission $LOGGHG$ and other firm fundamentals as before. We include the year fixed effect denoted as δ_t and the industry fixed effect denoted as μ_j in the model. $\varepsilon_{i,t}$ is the residual of the model. All standard errors are clustered at the industry level.

$$INCLUSION_{i,t} = \alpha + \beta LOGGHG_{i,t} + \gamma' X_{i,t} + \mu_j + \delta_t + \varepsilon_{i,t} \tag{3}$$

We present regression results in table 9, where dependent variables in the first three columns are the number of inclusions and the probability of inclusion into ESG-related funds in the other three columns. In columns 3 and 6, we add other control variables of institutional ownership into the regressions.

[Insert Table 9 near here]

As seen from table 9, the higher the carbon emission, the less likely it would enter an ESG-related fund's portfolio. In the second column, the regression coefficient is -0.1326 with a t-statistic of -2.98, which suggests that the more carbon-intensive a firm is, the less likely its stock would be included in an ESG-related fund. In columns 4 to 6, regression results are also economically significant, where the regression coefficient is -0.0036 in the fifth column, with a t-statistic of -1.94. Interestingly, the coefficients in front of the carbon emissions become larger and more significant when we control for the year-fixed effect in the second and fifth columns.

We examine the determinants and compare carbon emissions estimated by the XGBoost algorithm and the original emission data provided by the Trucost database in table 10. Following Bolton and Kacperczyk (2021a), we regress three different measures of carbon emission as equation 4 shows. Regression results suggest the determinants are pretty similar for carbon emissions and emission intensities, whereas emission growth rates differ significantly.

$$GHG_{i,t} = \gamma' X_{i,t} + \mu_j + \delta_t + \varepsilon_{i,t} \tag{4}$$

The dependent variables include the logarithmic value of carbon emission, emission growth rate, and emission intensity defined as carbon emission over firm sales, on firm fundamentals. The independent variables $X_{i,t}$ include a host of financial characteristics like firm size, book-to-market ratio, leverage ratio, investment ratio, ROE, HHI index, Plant, property & equipment, sales growth rate, and EPS growth rate as before. μ_j denotes industry fixed effects, δ_t denotes year fixed effects, and $\varepsilon_{i,t}$ is the residual.

[Insert Table 10 near here]

In table 10, the first and second columns report regression results where the dependent variable is the logarithmic value of carbon emissions. The first column relies on the original data set provided by the Trucost database, and the second column uses the full sample estimated with the XGBoost algorithm. Regression coefficients in columns 1 and 2 are largely similar, as the statistical significance and economic magnitude are largely the same. The coefficient in front of size is 0.5221 and 0.2235 for both samples, with t-statistics 19.34 and 5.33, respectively. Only the ROE and HHI index coefficients yield different results. The reason behind this difference may be because Trucost firms are larger and more profitable ones which could potentially bias the results, whereas the whole sample estimated by the XGBoost contains smaller firms. In columns 3 and 4, we use emission growth rate as the dependent variable and columns 5 and 6 report regression results with the emission intensity variable. Similarly, columns 3 and 5 use the Trucost sample, and columns 4 and 6 use the XGBoost sample. Overall, regression results are largely the same, suggesting that the estimated sample by our machine learning algorithm captures the emission pattern that can be explained by firm fundamentals. The only difference lies in the emission intensity variable, as the magnitude is much higher for the estimated sample as compared to the Trucost sample. We argue this is because our estimated data has comparatively produced higher emission estimates, and the sales are lower, which induces a much higher intensity estimation. In untabulated results, the mean value of emission intensity for the estimated sample and the Trucost samples are 49.06 and 1.84, and the medians are 2.81 and 1.84.

We rely not only on empirical design and economic analysis to validate our data set, but we also follow the traditional Machine learning approach method to show our model's robustness and prediction. We report a training result comparison with linear models using different training sets and cross-validate model parameters. We set different learning rates and tree depths of the model. We display training results in the appendices. Our model is robust under various tests ranging from empirical economic analyses to mainstream machine learning tests, which make convincing predictions and yield reasonable results. We believe the data set is valid and a good complement to the existing data set estimated by major data vendors. This data set can surely be used for many analyses of carbon emissions in the US stock market.

4. Empirical results on carbon emission and asset pricing

4.1. How is carbon premium priced in the cross-section of stock returns?

We examine the link between carbon emission and cross-sectional stock returns from 2002 to 2021 for firms listed in three major US stock markets. We explore the cross-sectional properties of stock returns with firms' carbon emissions, and we further examine the relation between (low-)carbon premia with common risk factors. To examine whether and how carbon emission has been priced in the stock markets over the past two decades, we first follow the pooled OLS regression model used in Bolton and Kacperczyk (2021a) as follows:

$$RET_{i,t} = \alpha + \beta GHG_{i,t} + \gamma' X_{i,t-1} + \delta_t + \mu_j + \varepsilon_{i,t}, \tag{5}$$

where the dependent variable is the stock return of firm i in year-month t, and the generic independent variable $GHG_{i,t}$ of interest are three measures of firms' carbon emissions, including the logarithmic value of carbon emission, emission growth rate, and emission intensity defined as carbon emission over firm sales. We use three different samples for empirical estimation. The first sample is the dataset obtained from the Trucost database, with a sample period from 2002 to the start of 2016, and has 172059 observations. The second sample is emission data predicted from the XGBoost model with the same sample period from 2002 to 2016. It has 533001 observations. The third sample period focuses on the period from the start of 2016 to 2021. We chose the start of 2016 as the Paris Agreement was signed at the end of 2015, and institutional investors began to fully recognize the notion of climate change. This set of data has 231149 observations. Other financial variables $X_{i,t-1}$ include firm size, book-to-market ratio, leverage ratio, momentum, investment ratio, ROE, HHI index, Plant, property & equipment, Beta, return volatility, sales growth rate, and EPS growth rate. We also include year-month fixed effects δ_t and include industry-fixed effects μ_j separately in the regressions, and we cluster standard errors at the 2-digit GIC industry and year levels.

We report regression results in table 11. In the first to the sixth columns, we replicate the estimation in Bolton and Kacperczyk (2021a), where we either report regression results with or without industry-fixed effects. In the first column, where we regress stock returns on the logarithmic value of carbon emissions, the coefficient is 0.0337 with a t-statistic of 1.44, suggesting that higher-emission firms earn higher stock returns. However, the result is not significant. Similar to Bolton and Kacperczyk (2021a), regression results are more pronounced after we control for industry fixed effects, with both more economically and statistically higher significance. Apart from the baseline estimation

with carbon emission, emission growth rates are highly significant, and emission growth rates are not. In the next set of samples in panel B, the positive carbon premium becomes negatively insignificant, as the economic magnitude shrinks to -0.0077 with a t-statistic of -0.83. This could be a result of the inclusion of more small-cap firms that do not disclose carbon emissions.

In the last sample, where we narrow down the observation period into 2016 to 2021 with data estimated by XGBoost, the baseline estimation yields completely different results than the positive relationship in the first two panels. The regression coefficients are -0.0798 and -0.0606, with t-statistics of -2.39 and -1.81, respectively. This result implies a shift in investors' preference for low-emission stocks. More interestingly, the significance is less pronounced once we control for the industry-fixed effect, which may suggest that investors invest in low-carbon stocks on an industry basis. Besides, the emission intensity and return relationship remain significantly negative in columns 17 and 18. In the appendices, we show that this reversed emission-return relationship is also prominent after the Paris Agreement with the Trucost data.

Overall, empirical results show that in more recent years, the higher the carbon emission a firm had, the lower realized returns it would earn. In unreported regressions, we estimate the low-carbon return premium using carbon emission intensity normalized with different firm fundamentals, including total assets, non-current assets, Plant, property & equipment, and firm's market capitalization, and the negative relationship is consistent. Note that our estimation performs poorly in terms of emission growth rates, as we are performing emission estimation on a cross-sectional basis resulting in large deviations in the data.

[Insert Table 11 near here]

To examine the time-varying emission-return relationship, we perform regression in 5 by year with three different emission measures. We control for industry-fixed effects each year and report regression coefficients as well as t-statistics in front of carbon emission. The sample period is from 2002 to 2021. Regression results are displayed in table 12. In the first set of rows, we report baseline regression results where the emission measure is logarithmic carbon emission. The signs of the regression coefficients were indefinite before 2015, and they turned consistently negative from 2016 to 2020. The average carbon premium during this period is -0.0019 with a t-statistic of -0.15 for the 2002-2011 period and -0.0431 with a t-statistic of -1.88 for the 2012-2021 period. The baseline results imply that there appears to be a time-varying preference for low-carbon assets as investors divert their portfolio from brown firms and tilt more towards green firms. This trend seems to have begun after 2015. The negative emission-return relationship was more pronounced from 2016 to 2020 and was highest in the year 2020, during which the notion of responsible investment gained increasing recognition from the industry. As for the second set of

regressions where the dependent variable is emission growth rates, regression results yield inconsistent and ambiguous results, as the coefficients are not consistently negative. This can be attributed to volatile estimation across periods due to our estimation methodology. For the third set of regressions, where the dependent variable is emission intensity, the negative relationship is more pronounced and is significant even before 2015.

[Insert Table 12 near here]

For more illustrative plots, we make a comparison between the cumulative carbon premia estimated with the original Trucost data and the XGBoost predicted data. We adjust the magnitudes in terms of the unit standard deviation of the logarithmic emission at each cross-section following Bolton and Kacperczyk (2021a). Subfigure A of figure 5 shows that prior to the Paris Agreement, there was a strong upward trend in the Trucost sample, and the positive premium is insignificant for the sample estimated by XGBoost. Since the Trucost data is limited in the number of firms, there appears to be a drastic increase in carbon premium around 2008, while it is relatively stable for the XGBoost estimated premium. After the Paris Agreement, both sets of premia appear to be significantly negative, and the negative premium is much more pronounced for the XGBoost estimated sample. The result suggests that investors are diverting their portfolios towards low-carbon stocks at a speed way faster than previous researchers could have imagined, and this result is consistent with the findings in van der Beck (2021).

In subfigure B of figure 5 where we include industry-fixed effects in the regressions, the cumulative premia is more contrasting for the whole sample estimated with XGBoost and Trucost data. First of all, the sharp contrast before and after the Paris Agreement is less significant for the Trucost data, as the emission premium keeps increasing after 2004. The result echoes with the findings in Bolton and Kacperczyk (2021a). However, adding industry-fixed effects makes the results confusing after 2016. As for the XGBoost estimated data, the carbon premium before the Paris Agreement is less pronounced and is slowly decreasing. After the Paris Agreement, the carbon premia drops significantly just as the results in panel A.

[Insert Figure 5 near here]

In figure 6, we present more robust evidence supporting the low-carbon premium hypothesis after the Paris Agreement with the data sample estimated with XGBoost. In subfigure A, we first sort firms based on their current year's carbon emission based on the emission data estimated by the XGBoost algorithm into five quintiles, and we form either value-weighted or equal-weighted portfolios. The sample period is from the start of 2002 to the end of 2021. We report the high-minus-low portfolio returns in this figure and their summary statistics in the appendices. In figure 6, the portfolio returns appear to

be positive before 2012 for value-weighted returns in a short period, and the relationship soon reversed afterward. This negative emission-return relationship became increasingly more significant after the end of 2015 and peaked at the end of 2020. The results suggest that the carbon premium did exist before the Paris Agreement, but investors' preferences significantly changed in the last decade. In subfigure B, the hi-minus-low portfolio results also produce negative cumulative returns after 2012. Overall, the results show that our sample, which has a broader coverage of listed firms, makes the positive carbon premium more significant by allowing researchers to apply more classic empirical approaches and further complementing the findings by (Bolton and Kacperczyk, 2021a).

[Insert Figure 6 near here]

4.2. A flow-based mechanism

As Pástor et al. (2022) and van der Beck (2021) show, the equilibrium price of the ESG-related assets is mostly driven by the responsible investors who value ESG assets, and therefore the positive realized ESG premium we are observing are essentially driven by institutional investors' money flows.

We complement their research by showing that this is the same potential mechanism that drives our result. We first start with a universe of institutional investors' holdings with the 13F filings from Thomson Reuters that document their holdings of both high-carbon firms and low-carbon firms. In the US, the regulation mandates that investment companies with AUM over $100M must report their respective holdings via quarterly SEC 13F filings. We compute the money flow for each stock following van der Beck (2021), which is very similar to the fund flow measure commonly used in the mutual fund literature. This measure is shown in equation 6,

$$FLOW_{i,t} = A_{i,t} - A_{i,t-1} \times (1 + RET_{i,t}) \tag{6}$$

In equation 6, $A_{i,t}$ denotes the total asset under management in year t by all eligible institutional investors for firm i, and this measure is averaged across all the quarters in each year since the 13F is filed at a quarterly frequency. The $RET_{i,t}$ is the annual return for firm i at the end of year t. Thus, the measure $FLOW_{i,t}$ calculates net purchases (or redemption) from all the major investment companies.

We provide reduced-form evidence in table 13 which shows that institutional investors are diverting their portfolio weights from the high-carbon firms to the low-carbon firms. To examine the flow-emission relationship, we regress the logarithmic value of investor flow on firms' carbon emission, a time dummy that indicates the time after 2016, and their interaction term. We also host a set of control variables similar to equation 4 and

control for industry fixed effect. We double-cluster the standard errors at the 2-digit GIC industry and year level.

[Insert Table 13 near here]

In table 13, the regression table shows that there is a negative relationship between firms' carbon emission and investor flow post the Paris Agreement. In the first and second columns, regression results are -0.0494 and -0.0492, with t-statistics of -5.55 and -4.96, respectively. However, for the LOGGHG variable itself, the regression results are significantly positive, which is quite intuitive as carbon emission is associated with firm size, and larger firm size induces higher investor flow.

Moreover, the regression results in columns 3 to 4 are also significant for coefficients in front of the interaction term, whereas the results are insignificant in columns 5 and 6. This is consistent with the observations from table 11, as the change in carbon premium is insignificant for carbon emission measured by emission intensities around the Paris Agreement. The intensity measure seems quite persistent.

In figure 7, we present more illustrative evidence showing investor flow for stocks of different emission quintiles. We first sort stocks of firms based on their logarithmic value of carbon emissions into 5 quintiles and then plot the cumulative investor flows for each quintile. The orange line represents the high-emission portfolio and the blue represents the low-emission portfolio. We plot the confidence intervals at 95% levels. As figure 7 suggests, the high-emission quintile consistently experiences higher investor flows as compared to the lower quintile group, and the relationship soon reversed after the Paris Agreement in 2016. In the years 2002 to 2010, the difference in investor flows is barely noticeable.

[Insert Figure 7 near here]

Next, we examine whether there is flow-induced carbon premium and the heterogeneity in emission-return relationship. We first sort firms of their logarithmic value of carbon emissions into 5 quintiles and then regress monthly returns on the investor flows. We estimate a pooled OLS similar to equation 5 and control for year-month fixed effect and industry fixed effect. Since previous research suggests that investor flow has strong predictive power for stock future returns (the smart money effect)(Lou, 2012), we control for lagged flows in the regression. We double cluster standard errors at the 2-digit GIC industry and year level and the regression results are shown in table 14.

[Insert Table 14 near here]

As table 14 suggests, there is a significant flow-induced carbon premium for the low-carbon portfolios. In panel A, columns 1 and 2 where the stocks of firms belong to the

lowest emission group, the regression coefficients are 1.4940 and 1.4321, with t-statistics of 4.31 and 4.33, respectively. In columns 3 and 4 the regression results are also significant. However, the positive flow-return relationship becomes insignificant. However, the results in the fourth quintile are also significant but the magnitude is much smaller. Overall, regression results suggest that this flow-driven return is only predominant among the low-carbon firms, as investors began to purchase green stocks which gave rise to positive price pressures. However, the investor flows out of the brown firms are not necessarily associated with the lower realized returns may be because of a small margin of hedge funds or investors who do not care about responsible investing purchase the brown stocks for investment returns, which leads to higher realized returns. In panel B and panel C where we rely on emission growth rate and emission intensity, the flow-induced stock return is also prominent among the low-carbon quintile.

4.3. The carbon premium and common risk factors

Finally, we examine the relationship between carbon premia and risk factors. We estimate a time-series regression model using monthly premium, which is estimated from monthly cross-sectional regression in 5. We run the following regression:

$$RISKPRMM_t = \alpha + \beta' FACTOR_{i,t} + \varepsilon_t \tag{7}$$

where the dependent variable is the monthly risk premium estimated from equation 4, and we substitute the carbon emission with three different measures. $FACTOR_{i,t}$ denotes various common risk factors, including the market factor as the CAPM model, factors from other widely adopted models like the Fama-French three-factor and the five-factor model including SMB, HML, profitability factor RMW, investment factor CMA, a Betting-Against-Beta factor BAB that accounts for margin investments in (Frazzini and Pedersen, 2014), and Pastor-Stambaugh's liquidity factor (Pástor and Stambaugh, 2003). We also calculate the standard errors of the coefficients using the Newey-West robust estimator with 12 lags to adjust serial correlations. The main regression coefficient of interest is α, which measures the carbon premium after controlling for common risk factors. We use two different sample periods when estimating the alphas and risk loadings. The first sample is the whole data set estimated from the XGBoost model, which spans from 2002 to 2021. The second sample is the partial sample after the Paris Agreement from 2016 to 2021.

[Insert Table 15 near here]

Regression results in columns 1 and 2 of table 15 are significantly negative, suggesting that the risk premium is largely absorbed by common risk factors even though the risk premium is insignificant itself. However, the negative premium becomes even stronger

in columns 7 and 8 after we narrow down the sample period from 2016 to 2021. The estimated alphas in both columns are -0.0853 and -0.700 with and without controlling for common risk factors, with t-statistics of -5.16 and -2.97, respectively. In the next set of columns 9 and 10, we report risk premia estimated by emission growth rates, and the relationship is unsurprisingly insignificant. Finally, in columns 11 and 12, the negative premium is consistent in all two sample periods for risk premium estimated by carbon emission intensity. The carbon premium, as a unique risk component, cannot be solely explained by common risk factors like size and value. As an intrinsic risk factor that derives from investors' awareness of sustainability, it is largely independent.

5. Conclusion

In this paper, we adopt a novel method based on XGBoost to predict the carbon emissions of non-disclosure firms or firms that do not disclose carbon emissions to the public, especially before 2016. Under the hypothesis that similar firms produce similar carbon emissions, we use cosine similarity scores between firm pairs, the disclosure firms' carbon emissions and other firm fundamentals to predict carbon emissions for the non-disclosure firms. We estimate a large panel of carbon emissions based on this approach conduct asset pricing tests, and find that, on average, stocks of high-emission firms consistently underperform stocks of low-emission firms, especially after 2016. Common risk factors cannot explain the low-carbon risk premium after 2016, implying that there has been a positive shock to investors' ESG-related preference, and investors started to purchase more low-carbon stocks, which pushed up realized returns. We provide a flow-based mechanism to explain the negative carbon premium, especially after 2016, as (large) institutional investors are buying stocks of low-carbon firms and selling stocks of high-carbon firms after the Paris Agreement. We expect this phenomenon could be persistent in the next few years as responsible investors keep tilting more portfolio weights to the green stocks until in a new equilibrium, the emission-return relationship reverses to become positive.

In this paper, we also examine the fitness of our data set by designing a battery of tests with both empirical economic designs and machine learning experiments. Our data set proves robust and produces convincing results under different scenarios.

We believe our data set can be extensively used in the topic of climate finance, both in academic research and policy papers. Carbon emission is an important endogenous variable related to not only stock returns or corporate financial performance but also has a wider implication on the socioeconomic impact on its surroundings beyond finance. Our prediction method based on business similarity networks can also be adopted with other firm network data, to estimate other types of data sets including scope 2 and scope 3 data, comprehensive ESG ratings, corruption index, etc. Policymakers could also consider directly targeting heavy emission firms and punishing firms that misreport their

real carbon emissions with this data set.

References

Acharya, V. V., Johnson, T., Sundaresan, S., Tomunen, T., 2022. Is physical climate risk priced? evidence from regional variation in exposure to heat stress. Tech. rep., National Bureau of Economic Research.

Alekseev, G., Giglio, S., Maingi, Q., Selgrad, J., Stroebel, J., 2022. A quantity-based approach to constructing climate risk hedge portfolios. Tech. rep., National Bureau of Economic Research.

Ali, U., Hirshleifer, D., 2020. Shared analyst coverage: Unifying momentum spillover effects. Journal of Financial Economics 136, 649–675.

Aswani, J., Raghunandan, A., Rajgopal, S., 2022. Are carbon emissions associated with stock returns? Columbia Business School Research Paper Forthcoming .

Basu, S., Ma, X., Shen, M., 2023. The value of teamwork for firms' human capital. Working Paper .

Bernard, D., Blackburne, T., Thornock, J., 2020. Information flows among rivals and corporate investment. Journal of Financial Economics 136, 760–779.

Bolton, P., Halem, Z., Kacperczyk, M., 2022a. The financial cost of carbon. Journal of Applied Corporate Finance 34, 17–29.

Bolton, P., Kacperczyk, M., 2021a. Do investors care about carbon risk? Journal of financial economics 142, 517–549.

Bolton, P., Kacperczyk, M., 2021b. Global pricing of carbon-transition risk. Tech. rep., National Bureau of Economic Research.

Bolton, P., Kacperczyk, M., Samama, F., 2022b. Net-zero carbon portfolio alignment. Financial Analysts Journal 78, 19–33.

Bolton, P., Kacperczyk, M. T., 2020a. Carbon premium around the world .

Bolton, P., Kacperczyk, M. T., 2020b. Signaling through carbon disclosure. Available at SSRN 3755613.

Bolton, P., Kacperczyk, M. T., 2021c. Carbon disclosure and the cost of capital. Available at SSRN 3755613 .

Bolton, P., Kacperczyk, M. T., Wiedemann, M., 2022c. The co2 question: Technical progress and the climate crisis. Available at SSRN .

Bolton, P., Reichelstein, S., Kacperczyk, M. T., Leuz, C., Ormazabal, G., Schoenmaker, D., 2021. Mandatory corporate carbon disclosures and the path to net zero. Management and Business Review 1.

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

Bustamante, M. C., Frésard, L., 2021. Does firm investment respond to peers' investment? Management Science 67, 4703–4724.

Carhart, M. M., 1997. On persistence in mutual fund performance. The Journal of finance 52, 57–82.

Chang, D., Liao, G., Zheng, X., 2023. Private responsible engagements and esg performance. Working Paper .

Cheema-Fox, A., LaPerla, B. R., Serafeim, G., Turkington, D., Wang, H. S., 2021. Decarbonization factors. The Journal of Impact and ESG Investing .

Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785–794.

Choi, D., Gao, Z., Jiang, W., 2020a. Attention to global warming. The Review of Financial Studies 33, 1112–1145.

Choi, D., Gao, Z., Jiang, W., 2020b. Measuring the carbon exposure of institutional investors. The Journal of Alternative Investments 23, 12–23.

Choi, D., Gao, Z., Jiang, W., Zhang, H., 2022. Carbon stock devaluation. Available at SSRN 3589952 .

Cohen, L., Frazzini, A., 2008. Economic links and predictable returns. The Journal of Finance 63, 1977–2011.

Cohen, L., Gurun, U. G., Nguyen, Q. H., 2020a. The esg-innovation disconnect: Evidence from green patenting. Tech. rep., National Bureau of Economic Research.

Cohen, L., Malloy, C., Nguyen, Q., 2020b. Lazy prices. The Journal of Finance 75, 1371–1415.

Duan, T., Li, F. W., Wen, Q., 2021. Is carbon risk priced in the cross-section of corporate bond returns? Available at SSRN 3709572 .

Eisdorfer, A., Froot, K., Ozik, G., Sadka, R., 2022. Competition links and stock returns. The Review of Financial Studies 35, 4300–4340.

Engelberg, J., Gao, P., Parsons, C. A., 2012. Friends with money. Journal of Financial Economics 103, 169–188.

Engelberg, J., Gao, P., Parsons, C. A., 2013. The price of a ceo's rolodex. The Review of Financial Studies 26, 79–114.

Erel, I., Stern, L. H., Tan, C., Weisbach, M. S., 2021. Selecting directors using machine learning. The Review of Financial Studies 34, 3226–3264.

Frazzini, A., Pedersen, L. H., 2014. Betting against beta. Journal of financial economics 111, 1–25.

Garvey, G. T., Iyer, M., Nash, J., 2018. Carbon footprint and productivity: does the "e" in esg capture efficiency as well as environment. J Invest Manag 16, 59–69.

Gibson, R., Krueger, P., Mitali, S. F., 2020. The sustainability footprint of institutional investors: Esg driven price pressure and performance. Swiss Finance Institute Research Paper .

Glossner, S., 2021. Repeat offenders: Esg incident recidivism and investor underreaction. Available at SSRN 3004689 .

Görgen, M., Jacob, A., Nerlinger, M., Riordan, R., Rohleder, M., Wilkens, M., 2020. Carbon risk. Available at SSRN 2930897 .

Han, Y., Gopal, A., Ouyang, L., Key, A., 2021. Estimation of corporate greenhouse gas emissions via machine learning. arXiv preprint arXiv:2109.04318 .

Hege, U., Li, K., Zhang, Y., 2023. Climate innovation and carbon emissions: Evidence from supply chain networks. Available at SSRN 4557447 .

Hoberg, G., Phillips, G., 2010. Dynamic text-based industry classifications and endogenous product differentiation. Unpublished working paper, University of Maryland, College Park, MD .

Hoberg, G., Phillips, G., 2016. Text-based network industries and endogenous product differentiation. Journal of Political Economy 124, 1423–1465.

Hoberg, G., Phillips, G. M., 2018. Text-based industry momentum. Journal of Financial and Quantitative Analysis 53, 2355–2388.

Hsu, P.-H., Li, K., Tsou, C.-Y., 2022. The pollution premium. Journal of Finance, Forthcoming .

Ilhan, E., Sautner, Z., Vilkov, G., 2021. Carbon tail risk. The Review of Financial Studies 34, 1540–1571.

In, S. Y., Park, K. Y., Monk, A., 2017. Is "being green" rewarded in the market? an empirical investigation of decarbonization risk and stock returns. International Association for Energy Economics (Singapore Issue) 46.

Javadi, P., Yeganeh, B., Abbasi, M., Alipourmohajer, S., 2021. Energy assessment and greenhouse gas predictions in the automotive manufacturing industry in iran. Sustainable Production and Consumption 26, 316–330.

Jin, Z., Li, F. W., 2020. Geographic links and predictable returns. Available at SSRN 3617417 .

Lee, C. M., Sun, S. T., Wang, R., Zhang, R., 2019. Technological links and predictable returns. Journal of Financial Economics 132, 76–96.

Leippold, M., Yu, T., 2023. The green innovation premium: Evidence from us patents and the stock market. Swiss Finance Institute Research Paper .

Li, F., Lundholm, R., Minnis, M., 2013. A measure of competition based on 10-k filings. Journal of Accounting Research 51, 399–436.

Li, F., Zheng, X., 2024. Carbon awareness and return comovement. Working paper .

Lou, D., 2012. A flow-based explanation for return predictability. The Review of Financial Studies 25, 3457–3489.

Luccioni, A., Palacios, H., 2019. Using natural language processing to analyze financial climate disclosures. In: Proceedings of the 36th International Conference on Machine Learning, Long Beach, California.

Mardani, A., Liao, H., Nilashi, M., Alrasheedi, M., Cavallaro, F., 2020. A multi-stage method to predict carbon dioxide emissions using dimensionality reduction, clustering, and machine learning techniques. Journal of Cleaner Production 275, 122942.

Matsumura, E. M., Prakash, R., Vera-Munoz, S. C., 2014. Firm-value effects of carbon emissions and carbon disclosures. The accounting review 89, 695–724.

Matsumura, E. M., Prakash, R., Vera-Muñoz, S. C., 2022. Climate-risk materiality and firm risk. Review of Accounting Studies pp. 1–42.

Medina, P. C., Pagel, M., 2021. Does saving cause borrowing? Tech. rep., National Bureau of Economic Research.

Monasterolo, I., De Angelis, L., 2020. Blind to carbon risk? an analysis of stock market reaction to the paris agreement. Ecological Economics 170, 106571.

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

Pástor, L., Stambaugh, R. F., 2003. Liquidity risk and expected stock returns. Journal of Political economy 111, 642–685.

Pástor, L., Stambaugh, R. F., Taylor, L. A., 2021. Sustainable investing in equilibrium. Journal of Financial Economics 142, 550–571.

Pástor, L., Stambaugh, R. F., Taylor, L. A., 2022. Dissecting green returns. Journal of Financial Economics 146, 403–424.

Pedersen, L. H., Fitzgibbons, S., Pomorski, L., 2021. Responsible investing: The esg-efficient frontier. Journal of Financial Economics 142, 572–597.

Peng, L., Titman, S., Yönaç, M., Zhou, D., 2022. Social ties, comovements, and predictable returns. Comovements, and Predictable Returns (July 29, 2022) .

Rossi, A. G., Utkus, S. P., 2020. Who benefits from robo-advising? evidence from machine learning. Evidence from Machine Learning (March 10, 2020) .

Skiadopoulos, G., Faccini, R., Matin, R., 2023. Dissecting climate risks: Are they reflected in stock prices? Journal of Banking and Finance .

Stoffman, N., Woeppel, M., Yavuz, M. D., 2022. Small innovators: No risk, no return. Journal of Accounting and Economics 74, 101492.

Tantri, P., 2021. Fintech for the poor: Financial intermediation without discrimination. Review of Finance 25, 561–593.

Teng, H. W., Li, Y.-H., Chang, S.-W., 2020. Machine learning in empirical asset pricing models. In: 2020 International Conference on Pervasive Artificial Intelligence (ICPAI), IEEE, pp. 123–129.

Tomar, S., 2022. Greenhouse gas disclosure and emissions benchmarking. SMU Cox School of Business Research Paper .

van der Beck, P., 2021. Flow-driven esg returns. Swiss Finance Institute Research Paper .

Webersinke, N., Kraus, M., Bingler, J. A., Leippold, M., 2021. Climatebert: A pretrained language model for climate-related text. arXiv preprint arXiv:2110.12010 .

Zhang, S., 2023. Carbon returns across the globe. Available at SSRN 4378464 .

Zheng, X., 2022. How can innovation screening be improved? a machine learning analysis with economic consequences for firm performance. A Machine Learning Analysis With Economic Consequences for Firm Performance (February 28, 2022) .

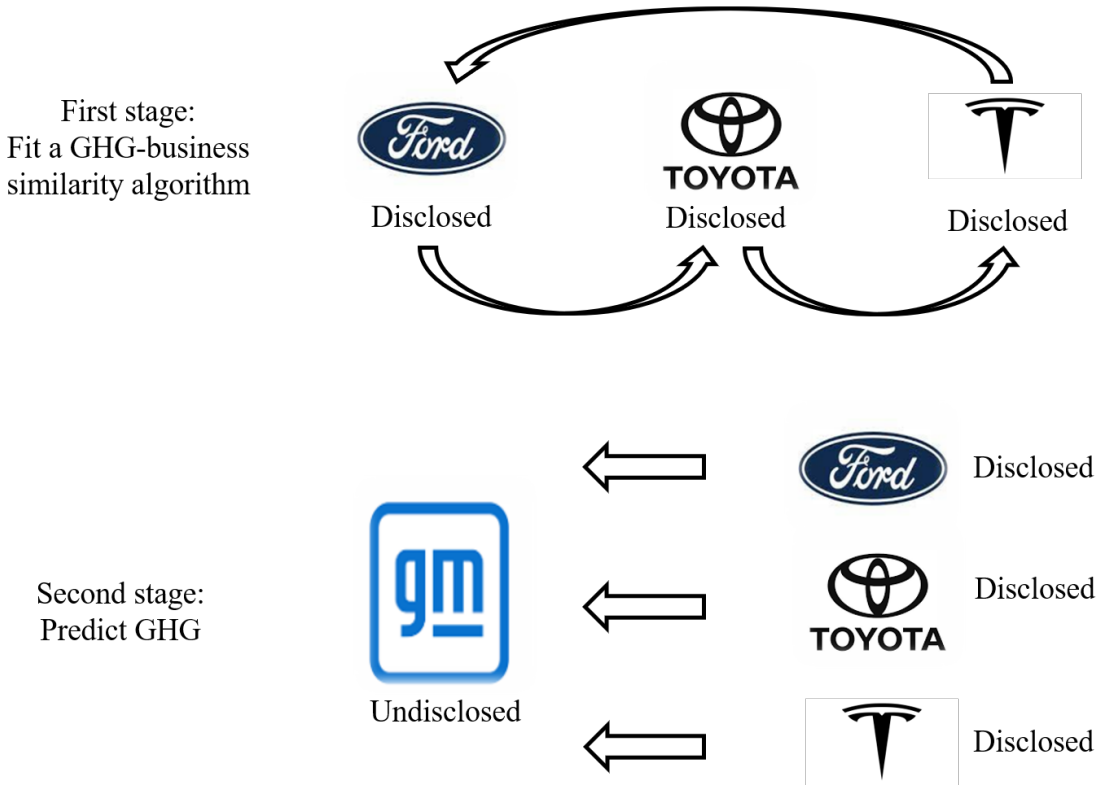


Fig. 1. Illustrative example of the prediction method. This figure gives a simple example of how we predict the carbon emissions of undisclosed firms. The first stage shows how we train an emission-business similarity algorithm for firms that disclose carbon emissions within the automobile industry. Then, in the second stage, we use the trained algorithm and use the carbon emissions of disclosed firms to predict the carbon emissions of the undisclosed firm.

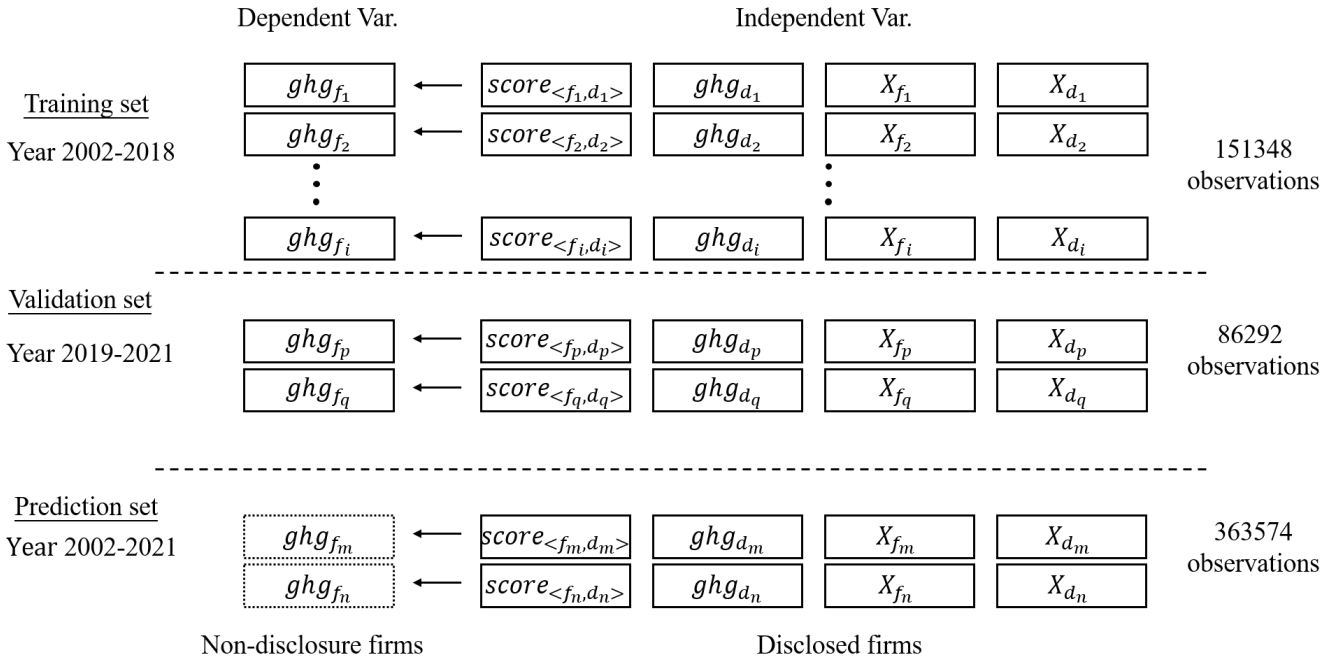
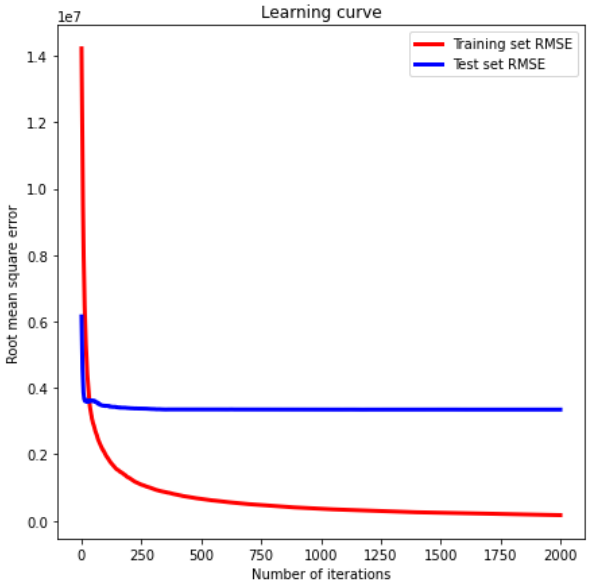
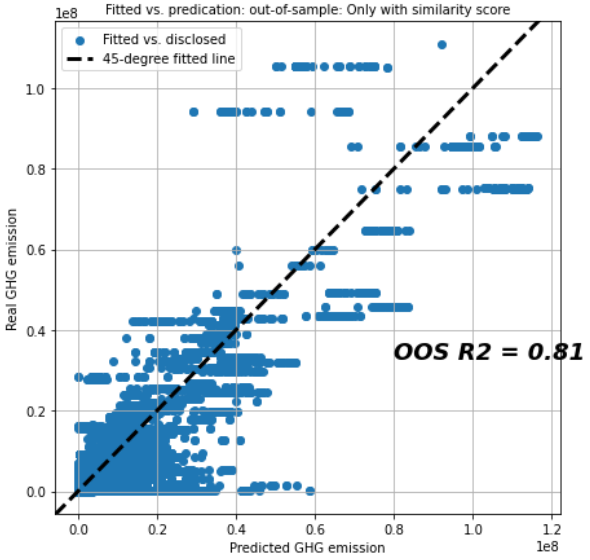


Fig. 2. Training methodology overview. This figure shows the partition for the training and validation set used for XGBoost learning. For similar firms that have observed carbon emission, or GHG (Greenhouse Gas) in our model, and other firm fundamentals, we use the carbon emission of disclosed firms ghg_d, their similarity $score_{<f,d>}$, firm fundamentals X_f or X_d (including a firm-fixed dummy for the non-disclosure firm f) to predict carbon emission of the non-disclosure firm ghg_f with XGBoost regression trees. The final sample we use has 229396 observations, where 145576 of them are used in the training set and 83820 observations are used in the test set. The training and test sets are obtained with firm pairs that have disclosed GHG on both sides, we duplicate and switch the firms to obtain a symmetric data panel, and we set observations from the year 2002 to 2018 as the training set and set observations from the year 2019 to 2021 as the test set. Finally, the prediction set is where we use the GHG of disclosure firms (on the right) to predict the carbon emission of non-disclosure firms. If a firm has multiple similar firms which leads to multiple predictions of carbon emissions ghg_f, then we compute the mean predicted emissions by averaging each predicted data.

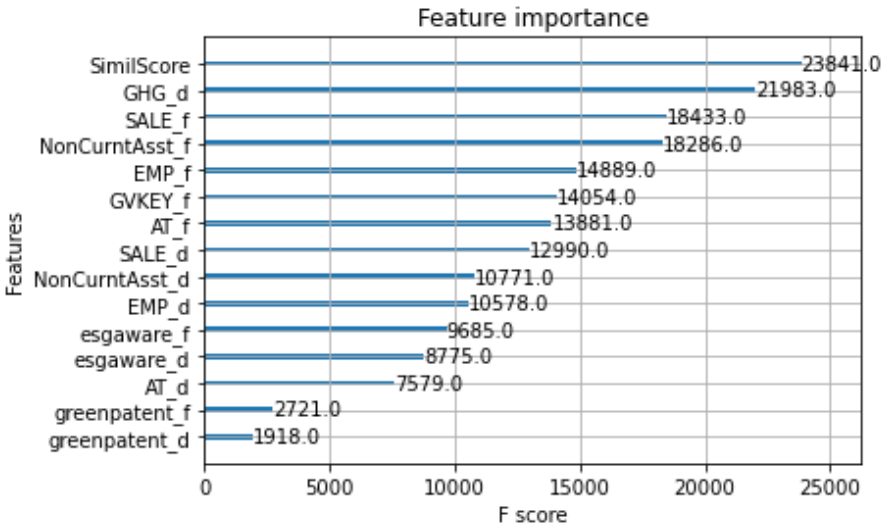


Subfigure A: XGBoost learning curve

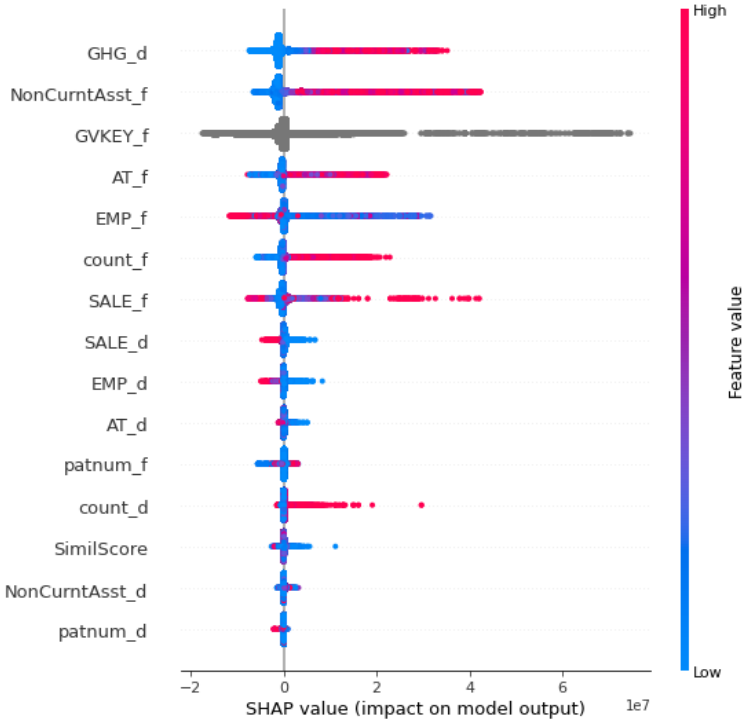


Subfigure B: Out-of-sample validation test

Fig. 3. XGBoost performance results. In subfigure A, we report the Root Mean Squared Error curve for both the training set and the validating set for the XGBoost model after 2 thousand times of iterations, where the blue line denotes the test curve and the red line denotes the training curve. In subfigure B, we report the out-of-sample validation tests. We use models trained from in-sample data to predict out-of-sample carbon emissions and compare the predicted out-of-sample values with real out-of-sample values. The x-axis indicates predicted GHG (scope 1 greenhouse gas emission), and the y-axis indicates real carbon emissions disclosed by firms or computed by the Trucost database. We add a 45-degree line to illustrate the fitness of our model.

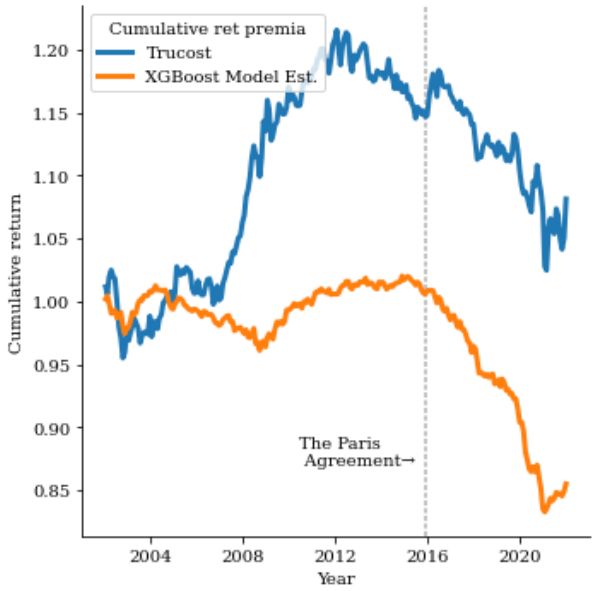


Subfigure A: Importance plot

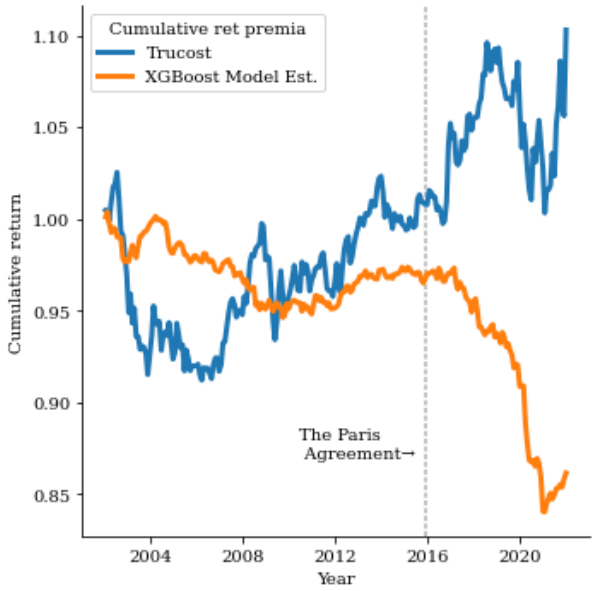


Subfigure B: SHAP value plot

Fig. 4. Variable importance contribution plot. We plot both the importance plot and the SHAP value plot for variables trained in the XGBoost model. In subfigure A, we illustrate the importance of each variable identified by the machine learning algorithm. The importance is measured by each feature's percentage of total predictive power on the x-axis. The name of each feature is on the y-axis, where the most important four features are similarity scores between two firms, the carbon emission of the disclosed firm, non-current assets for the target firm, and sales for the non-disclosure firm. The higher the feature importance, the stronger predictive power the variable has. In subfigure B, We present the SHAP value of each variable in the XGBoost model, which is a unified approach to explain the output in most tree models. The values in the x-axis show predictive power with positive or negative directions. Each dot represents an observation within the model. Higher inputs tend to have a higher SHAP value; a higher SHAP value means more importance or contribution to the model. All variables are displayed sequentially by their importance from top to bottom.

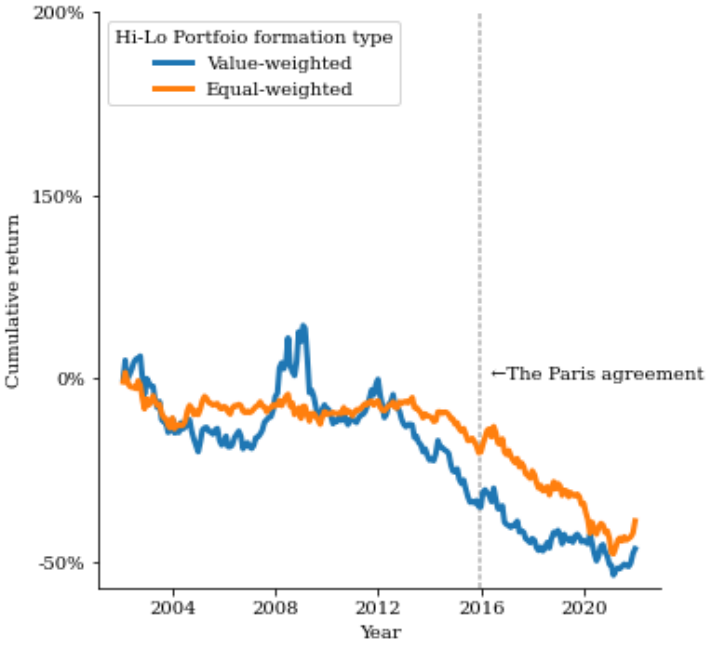


Subfigure A: Cumulative premia

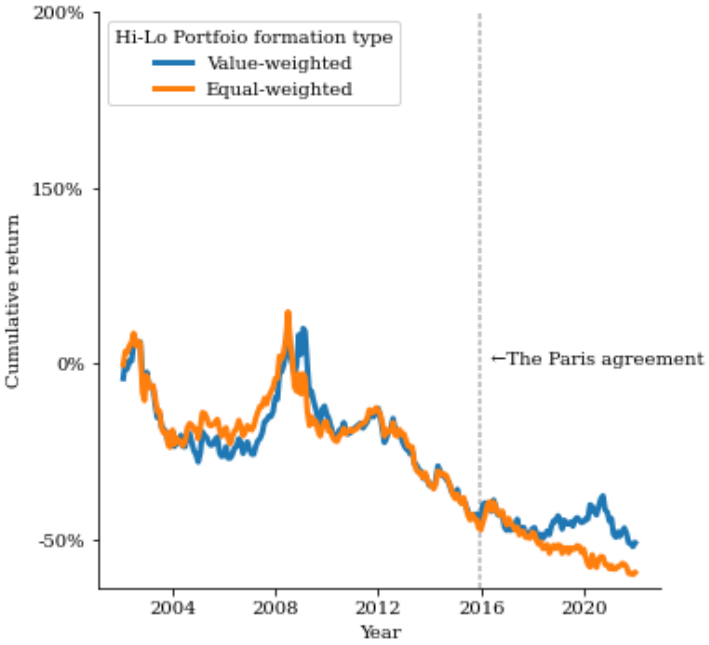


Subfigure B: Cumulative premia with industry FE

Fig. 5. Carbon cumulative return premia. This figure plots the cumulative return premia estimated from the cross-sectional regressions of monthly returns from 5. The independent variable of interest is the natural logarithmic value of carbon emission, and the dependent variable is monthly stock returns. We adjust the magnitudes in terms of the unit standard deviation of the logarithmic emission at each cross-section following Bolton and Kacperczyk (2021a). We use either the Trucost sample or the one estimated by the XGBoost algorithm for estimation, and the sample period is from 2002 to 2021. In subfigure A, we plot the cumulative premia without industry fixed effect, whereas we include industry fixed effect in subfigure B. The first vertical dashed line denotes the ratification of the Paris Agreement. The second vertical dashed line denotes the beginning of the year 2021. For the Trucost sample, there are 374 distinct firms in 2021, as a result, we do not report cumulative return premia after 2021 for the blue line. For the XGBoost sample, there are 4041 distinct firms in 2021.



Subfigure A: XGBoost data sample



Subfigure B: Trucost data sample

Fig. 6. Cumulative returns for high-minus-low carbon emission portfolios. This table plots cumulative returns for value-weighted or equal-weighted hi-lo portfolios sorted by logarithmic scope 1 carbon emissions at year t, where the blue line is the cumulative return of value-weighted portfolios, and the yellow line is the equal-weighted portfolio return. In subfigure A, we estimate the cumulative returns with the XGBoost predicted dataset, and in subfigure B, we use the original dataset provided by Trucost. The time period is 2002 to 2021. The vertical dashed line denotes the ratification of the Paris Agreement.

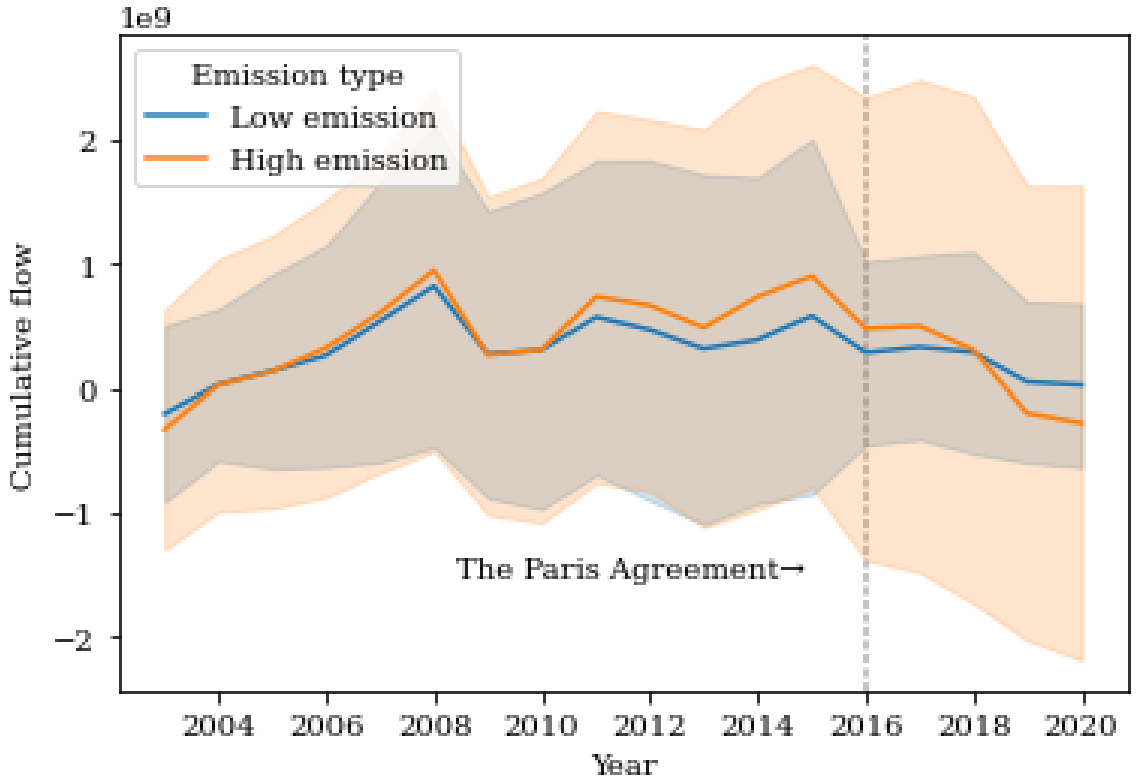


Fig. 7. Investor flow for stocks of firms with different levels of carbon emissions. This figure shows flow dynamics into low-carbon stocks and high-carbon stocks. We first sort firms into five quintiles based on their carbon emissions and calculate cumulative investor flows for each portfolio. The orange line represents the high-emission portfolio and the blue represents the low-emission portfolio. We plot the confidence intervals at 95% level. The vertical dashed line denotes the ratification of the Paris Agreement at the end of the year 2015.

Table 1: Firm similarity score by year

	N	Mean	Std	Min	25%	50%	75%	Max
2002	1064	0.07	0.10	0.00	0.02	0.05	0.09	0.82
2003	1984	0.07	0.08	0.00	0.02	0.06	0.09	0.82
2004	2636	0.07	0.07	0.00	0.03	0.06	0.09	0.82
2005	3962	0.06	0.06	0.00	0.02	0.05	0.09	0.82
2006	3758	0.06	0.07	0.00	0.02	0.05	0.09	0.83
2007	3902	0.06	0.07	0.00	0.02	0.05	0.09	0.82
2008	4220	0.06	0.07	0.00	0.02	0.05	0.09	0.81
2009	4398	0.06	0.06	0.00	0.02	0.05	0.09	0.80
2010	4498	0.06	0.06	0.00	0.02	0.05	0.09	0.80
2011	4908	0.06	0.07	0.00	0.02	0.05	0.09	0.81
2012	5240	0.06	0.06	0.00	0.02	0.05	0.09	0.80
2013	5984	0.07	0.07	0.00	0.02	0.05	0.09	0.81
2014	6132	0.06	0.06	0.00	0.02	0.05	0.09	0.81
2015	6490	0.06	0.06	0.00	0.02	0.05	0.09	0.81
2016	27640	0.07	0.07	0.00	0.02	0.05	0.09	0.81
2017	30288	0.07	0.07	0.00	0.02	0.05	0.09	0.85
2018	34244	0.07	0.07	0.00	0.02	0.05	0.09	0.84
2019	37342	0.07	0.07	0.00	0.02	0.05	0.09	0.80
2020	42494	0.07	0.07	0.00	0.02	0.05	0.09	0.73
2021	6456	0.07	0.07	0.00	0.02	0.05	0.09	0.66

This table reports firm cosine similarity pairs from 2002 to 2021. Similarity data is obtained from Hoberg and Philips data library. For each firm, we keep its top 20 most similar firm pairs and include them in the dataset. We report firm similarity scores by their mean, standard deviation, and other quintile summary statistics.

Table 2: Summary statistics of the machine learning sample

	N	Mean	Std	Min	25%	50%	75%	Max
Similarity Score	237640	0.07	0.07	0.00	0.02	0.05	0.09	0.85
LOGGHG	237640	10.58	3.30	0.00	8.47	10.39	12.61	18.92
LOGSALE	237640	7.35	2.21	0.00	6.18	7.61	8.91	13.23
LOGAT	237640	7.94	1.92	1.17	6.59	7.94	9.33	13.22
LOGNCT	237640	7.24	2.39	0.00	5.80	7.43	8.94	13.12
LOGEMP	237640	8.39	2.11	0.00	7.10	8.61	9.83	14.65
ESGAWARENESS	237640	6.75	12.02	0.00	0.00	2.00	8.00	137.00
GREENPATENT	237640	3.79	25.89	0.00	0.00	0.00	0.00	848.00

This table reports summary statistics of our dataset in three different panels. In our sample, only 237640 firm similarity score pairs have carbon emissions on both sides. We include the logarithmic value of firm sales, total assets, and non-current assets, which is defined by subtracting current assets from firms' total assets, and the number of employees into the machine learning algorithm. Additionally, we include two unique features in the algorithm. The first feature is firms' carbon awareness, which is a text-based measure computed from firms' 10-K and represents firms' desire to promote a sustainable business. The second feature is the number of green patents firms possess. The sample period is from 2002 to 2021.

Table 3: Number of disclosed firms in the dataset

	Number of disclosed firms in the dataset				
Year	Trucost	Thomson Reuters	CDP	Aswani et al. (2022)	XGBoost Estimated
2002	629	4	4		2952
2003	851	10	1		3703
2004	1026	20	12		4073
2005	1260	50	37	700	4406
2006	1275	171	54	706	4440
2007	1237	309	77	693	4367
2008	1251	367	105	690	4328
2009	1265	500	329	709	4162
2010	1258	550	564	704	4011
2011	1252	588	801	715	3938
2012	1252	597	900	727	3909
2013	1350	572	998	800	3946
2014	1372	585	1217	829	4049
2015	1377	659	1135	859	4117
2016	3265	722	1472	2369	4281
2017	3286	797	1469	2509	4228
2018	3363	895	1501	2645	4242
2019	3393	1066	1418	1992	4279
2020	3154	1110	1436		4329
2021	385	398	356		4453
Average	1675	499	694	1176	4111

In this table, we report the number of listed firms that have disclosed (or estimated by data vendors) available carbon emission data from different data vendors in columns 1 to 3. The databases include S&P Trucost, Thomson Reuters, and the Carbon Disclosure Project, which all began to provide data after 2002, but all with a very limited number of firms. We report the number of firms used in Aswani et al. (2022), which replicates Bolton and Kacperczyk (2021a) also using the Trucost database. In the last column, we report the number of firms with carbon emissions estimated by our method starting from 2002. The bottom line reports the average number of firms that have disclosed carbon emissions in each source.

Table 4: Emission comparison between different data sets by year

	Panel A: Trucost data					Panel B: Xgboost estimated				
	Distinct firms	Mean	Std	Median	Year-mon obs	Distinct firms	Mean	Std	Median	Year-mon obs
2002	629	12.05	2.54	11.90	6863	2952	11.43	3.96	12.68	32228
2003	851	11.60	2.66	11.35	9594	3703	11.45	3.75	12.66	40864
2004	1026	11.58	2.66	11.34	11699	4073	11.43	3.73	12.66	44688
2005	1260	11.37	2.68	11.19	14318	4406	11.43	3.64	12.64	46909
2006	1275	11.40	2.67	11.22	14432	4440	11.59	3.49	12.73	47311
2007	1237	11.40	2.64	11.21	13890	4367	11.63	3.51	12.80	46606
2008	1251	11.43	2.62	11.25	14025	4328	11.59	3.55	12.76	46319
2009	1265	11.30	2.62	11.07	14407	4162	11.56	3.53	12.72	46534
2010	1258	11.35	2.61	11.16	14465	4011	11.61	3.49	12.76	45738
2011	1252	11.34	2.62	11.08	14462	3938	11.49	3.69	12.79	43870
2012	1252	11.31	2.64	11.10	14484	3909	11.57	3.61	12.77	43285
2013	1350	11.25	2.65	11.07	15423	3946	11.52	3.58	12.68	43339
2014	1372	11.23	2.69	11.01	15314	4049	11.41	3.66	12.62	43275
2015	1377	11.21	2.65	11.02	15423	4117	11.29	3.70	12.49	42995
2016	3265	9.49	2.95	9.44	35521	4281	10.31	3.30	10.87	45471
2017	3286	9.48	2.98	9.41	36401	4228	10.25	3.28	10.70	45991
2018	3363	9.46	3.01	9.42	36984	4242	10.11	3.30	10.50	45611
2019	3393	9.39	3.02	9.35	36947	4279	9.98	3.32	10.30	45666
2020	3154	9.09	3.01	8.95	34567	4329	9.60	3.37	9.85	46666
2021	385	7.61	2.37	7.57	4441	4453	10.45	3.74	11.68	47236

This table compares the estimated data set with the original data set obtained from the Trucost database in detail. For each data set, we report summary statistics of its scope 1 logarithmic greenhouse gas emission by year from 2002 to 2021. In the last columns in each panel, we also report the number of firm-month observations in data samples.

Table 5: Scope 1 carbon emissions by industry

Panel A: Industry emission and summary basic summary stat					
Two digit GIC name	Firm-year Observation	Mean	Std	Median	Distinct firms
Utilities	2204	14.18	2.25	14.72	168
Materials	4111	12.42	3.41	13.02	387
Energy	6073	12.18	3.76	12.93	677
Consumer Staples	2961	11.78	3.28	12.55	297
Consumer Discretionary	9666	11.43	2.80	12.25	1087
Industrials	9593	11.09	3.64	12.03	931
Information Technology	12768	11.09	3.19	12.38	1588
Communication Services	3391	10.99	2.88	11.45	379
Health Care	12312	10.79	3.40	12.28	1833
Financials	15817	10.00	4.25	11.91	1737
Real Estate	3240	9.58	3.72	9.94	265

This table presents carbon emissions by industry in detail. In panel A, We report 2-digit GIC industry classification emissions on the firm-year level. We sort industries based on average scope 1 firm logarithmic carbon emissions. In panel B, we report detailed carbon emissions and the number of firm-year observations in panel B with 6-digit GIC codes.

Table 5: Cont'd

Panel B: Detailed emission by six-digit industry classification					
Six digit GIC name	Firm-year Observation	Mean	Std	Median	Distinct firms
Multi-Utilities	452	15.15	1.47	15.92	31
Electric Utilities	819	15.03	1.46	15.88	58
Airlines	383	14.70	1.55	14.91	36
Independent Power and Renewable Electricity Producers	254	13.85	3.11	15.40	32
Gas Utilities	415	13.61	1.21	13.73	29
Construction Materials	218	13.50	2.24	13.46	18
Industrial Conglomerates	136	13.40	2.60	14.09	11
Air Freight & Logistics	272	13.12	2.62	13.35	25
Containers & Packaging	424	13.10	2.36	13.37	36
Road & Rail	716	12.70	2.80	13.03	65
Oil, Gas & Consumable Fuels	4697	12.42	3.79	13.11	529
Beverages	549	12.32	2.14	12.60	44
Metals & Mining	1668	12.28	3.49	12.85	163
Marine	231	12.27	3.17	13.18	26
Paper & Forest Products	264	12.26	3.91	13.34	24
Chemicals	1537	12.26	3.57	13.00	146
Food & Staples Retailing	552	12.12	2.99	12.90	66
Automobiles	230	11.93	3.37	12.95	28
Distributors	127	11.84	2.33	12.48	16
Health Care Providers & Services	2033	11.79	2.80	12.64	242
Food Products	1062	11.78	3.64	12.73	106
Household Products	202	11.77	3.36	12.16	13
Internet & Direct Marketing Retail	481	11.77	2.46	12.87	78
Specialty Retail	2090	11.72	2.40	12.58	210
Diversified Telecommunication Services	1001	11.66	2.58	12.32	122
Multiline Retail	326	11.59	2.38	11.84	29
Technology Hardware, Storage & Peripherals	831	11.35	3.31	12.55	101
Energy Equipment & Services	1376	11.33	3.51	12.41	148
Auto Components	647	11.31	2.82	11.89	64
Hotels, Restaurants & Leisure	2295	11.29	3.03	12.20	252
Building Products	504	11.28	3.22	11.96	52
Wireless Telecommunication Services	477	11.26	2.44	11.19	58
Textiles, Apparel & Luxury Goods	892	11.25	2.46	11.67	92
Health Care Technology	424	11.23	3.25	12.70	66
Trading Companies & Distributors	601	11.20	2.98	11.87	66
Semiconductors & Semiconductor Equipment	2569	11.19	2.72	12.15	243
Personal Products	423	11.18	2.86	12.26	53
Aerospace & Defense	1030	11.16	3.66	12.11	99
Diversified Financial Services	264	11.15	4.32	12.68	36
Water Utilities	264	11.09	2.53	11.36	18
Household Durables	1078	11.04	3.28	11.85	107
IT Services	1710	11.03	3.20	12.54	211
Pharmaceuticals	1988	11.02	3.22	12.29	310
Communications Equipment	1605	11.02	3.22	12.27	194
Diversified Consumer Services	626	10.99	2.49	11.01	75
Leisure Products	309	10.99	2.41	11.30	38
Entertainment	514	10.98	2.97	11.49	57
Electronic Equipment, Instruments & Components	2230	10.89	3.57	12.12	219
Software	3197	10.87	3.27	12.59	464
Construction & Engineering	601	10.84	3.63	11.74	56
Health Care Equipment & Supplies	2790	10.79	3.28	12.16	357
Thrifts & Mortgage Finance	2033	10.77	4.49	13.01	289
Life Sciences Tools & Services	757	10.62	3.24	11.68	87
Commercial Services & Supplies	1349	10.58	4.07	11.77	138
Capital Markets	2375	10.56	3.60	12.45	224
Machinery	1888	10.55	3.32	11.31	165
Media	976	10.52	3.03	10.59	81
Tobacco	173	10.41	4.76	12.09	15
Consumer Finance	591	10.39	3.58	12.18	68
Mortgage Real Estate Investment Trusts (REITs)	477	10.38	4.29	12.77	52
Professional Services	1008	10.37	3.60	10.93	89
Interactive Media & Services	423	10.24	3.14	10.47	61
Biotechnology	4320	10.21	3.71	12.26	771
Insurance	2485	9.83	3.61	10.09	236
Transportation Infrastructure	57	9.80	2.48	9.47	9
Equity Real Estate Investment Trusts (REITs)	2748	9.61	3.69	9.87	215
Electrical Equipment	817	9.60	4.13	11.14	94
Banks	7053	9.52	4.52	11.23	705
Real Estate Management & Development	492	9.38	3.91	10.28	50

Table 6: Summary statistics and variable correlations

Panel A: Summary stats

	N	Mean	Std	Min	25%	50%	75%	Max
Firm-yearl level observations								
LOGGHG	82213	11.07	3.62	0.00	9.83	12.33	13.31	15.92
GHGINTEN	80469	48.70	132.98	0.00	0.14	2.71	21.57	703.62
LOGSIZE	80955	13.36	2.05	8.53	11.90	13.41	14.83	17.19
LEVERAGE	82166	0.58	0.27	0.08	0.37	0.58	0.80	1.11
INVEST2A	81043	0.04	0.05	0.00	0.00	0.02	0.05	0.23
ROE	82015	0.00	0.42	-1.19	-0.05	0.08	0.15	1.21
HHI	82136	0.09	0.07	0.02	0.05	0.07	0.12	0.35
LOGPPE	78955	4.57	2.59	0.02	2.56	4.49	6.49	9.54
B2M	77172	1.05	1.95	0.06	0.31	0.57	0.95	11.90
SALESGR	76818	0.10	0.31	-0.54	-0.04	0.06	0.18	1.34
EPSGR	77864	-0.03	2.15	-8.50	-0.37	0.08	0.56	5.86
Firm-year-month level observations								
RETX	890602	1.02	16.47	-97.22	-5.79	0.43	6.63	1988.36
MOM	890522	1.12	4.72	-44.98	-1.01	0.90	2.88	169.02
VOLAT	890531	12.61	10.69	0.27	6.69	10.00	15.31	583.47
BETA	890602	1.23	1.08	-21.13	0.59	1.09	1.70	44.39

Panel B: Correlation matrix

	LOGGHG	GHGINTEN	LOGSIZE	LEVERAGE	INVEST2A	ROE	HHI	LOGPPE	B2M	SALESGR	EPSGR
LOGGHG	1.00										
GHGINTEN	0.20	1.00									
LOGSIZE	0.08	-0.41	1.00								
LEVERAGE	0.01	-0.10	0.07	1.00							
INVEST2A	0.13	-0.08	0.08	-0.09	1.00						
ROE	0.03	-0.23	0.27	0.14	0.05	1.00					
HHI	0.04	0.05	-0.03	-0.08	-0.03	-0.04	1.00				
LOGPPE	0.19	-0.42	0.69	0.23	0.39	0.23	-0.05	1.00			
B2M	0.06	-0.05	-0.18	0.08	0.04	0.00	-0.01	0.23	1.00		
SALESGR	0.00	-0.02	0.08	-0.06	0.06	0.05	0.00	-0.04	-0.07	1.00	
EPSGR	0.00	-0.05	0.15	-0.04	-0.02	0.26	-0.02	0.05	-0.06	0.19	1.00

This table presents summary statistics of all variables in Panel A and the correlation matrix in Panel B for the firm-year sample from 2002 to 2021. We report firm scope 1 carbon emissions and firm fundamentals for all listed US stocks. We report firm-year level variables, including the logarithmic value of emission, emission scaled by sales, firm size, leverage ratio, investment ratio, ROE, HHI index, Plant, property & equipment, Book-to-market ratio, sales, and EPS growth. We also report year-month level variables like return, momentum, volatility, and stock beta. In panel B, we report the Pearson correlation matrix in the lower triangle.

Table 7: State regulation and firm carbon emission

	LOGGHG			
	(1)	(2)	(3)	(4)
Regulated	-0.6162***	-1.0534***		
	(-8.62)	(-16.11)		
RegulateYears			-0.0458***	-0.1038***
			(-8.88)	(-18.23)
LOGSIZE	0.0031	-0.0043	0.0064	0.0239
	(0.09)	(-0.12)	(0.18)	(0.68)
B2M	0.1510***	0.1478***	0.1550***	0.1565***
	(3.70)	(3.68)	(3.77)	(3.91)
LEVERAGE	0.9404***	0.9862***	0.9676***	1.0747***
	(5.41)	(5.70)	(5.56)	(6.22)
INVEST2A	-2.1454***	-2.2888***	-2.0163***	-2.3921***
	(-2.89)	(-3.06)	(-2.71)	(-3.20)
ROE	0.0673	0.0888	0.0645	0.0657
	(0.84)	(1.12)	(0.80)	(0.83)
HHI	2.3538***	2.2605***	2.1045**	1.7431**
	(2.88)	(2.79)	(2.59)	(2.17)
PPE	0.0978***	0.1089***	0.1002***	0.0952***
	(2.92)	(3.25)	(2.99)	(2.85)
SALESGR	0.0750	0.0569	0.0752	0.0492
	(1.35)	(1.03)	(1.35)	(0.89)
EPSGR	0.0053	0.0058	0.0055	0.0060
	(0.73)	(0.81)	(0.76)	(0.83)
Const	T	T	T	T
Ind FE	T	T	T	T
State FE		T		T
R2	0.01	0.02	0.02	0.02
N	61739	61739	61739	61739

In this table, we examine the effect of states' regulation shock on firms' carbon emissions. We regress the firm's carbon emission intensity, which is defined as carbon emission scaled by firm sales on a dummy variable that indicates whether its state has announced a carbon emission target, or on a continuous variable that represents the number of years before or after the regulation shock. The control variables include sales, total assets, non-current assets, firm size, leverage ratio, book-to-market ratio, investment ratio, ROE, HHI index, Plant, property & equipment, sales growth rate, and EPS growth rate. We control for firm-or-industry-fixed effects and state-fixed effects in the regression. All standard errors are clustered at the firm level. The sample period is from 2002 to 2021.

Table 8: Transition matrix of firms in each emission quntiles

	Panel A: Transition Prob. after 1 year				
	Q1 L0	Q2 L0	Q3 L0	Q4 L0	Q5 L0
Q1 L1	**80.68%**	15.35%	3.38%	2.02%	1.14%
Q2 L1	10.88%	**67.60%**	17.27%	3.77%	1.69%
Q3 L1	3.22%	11.99%	**61.33%**	19.58%	4.48%
Q4 L1	3.18%	3.42%	14.70%	**61.26%**	15.95%
Q5 L1	2.05%	1.64%	3.32%	13.38%	**76.74%**
N	14443	14774	14447	14271	14875
	Panel B: Transition Prob. after 3 years				
	Q1 L0	Q2 L0	Q3 L0	Q4 L0	Q5 L0
Q1 L3	**68.46%**	25.68%	7.01%	3.15%	1.85%
Q2 L3	14.33%	**50.47%**	27.99%	8.84%	3.33%
Q3 L3	6.41%	13.79%	**42.83%**	28.95%	8.50%
Q4 L3	6.60%	6.46%	16.12%	**43.96%**	21.09%
Q5 L3	4.19%	3.61%	6.05%	15.10%	**65.22%**
N	11120	11705	11118	10754	12031
	Panel C: Transition Prob. after 5 years				
	Q1 L0	Q2 L0	Q3 L0	Q4 L0	Q5 L0
Q1 L5	**60.06%**	29.97%	13.51%	4.68%	2.07%
Q2 L5	15.36%	**40.80%**	29.97%	16.14%	5.17%
Q3 L5	8.26%	15.00%	**33.24%**	28.35%	12.19%
Q4 L5	10.02%	8.74%	15.89%	**35.06%**	21.87%
Q5 L5	6.30%	5.49%	7.40%	15.77%	**58.70%**
N	8513	9121	8586	8103	9579
	Panel D: Transition Prob. after 7 years				
	Q1 L0	Q2 L0	Q3 L0	Q4 L0	Q5 L0
Q1 L7	**53.66%**	34.01%	17.64%	5.34%	3.42%
Q2 L7	15.85%	**35.28%**	32.43%	19.89%	6.14%
Q3 L7	9.41%	14.62%	**27.16%**	28.98%	13.40%
Q4 L7	12.61%	9.33%	15.00%	**30.79%**	21.40%
Q5 L7	8.47%	6.76%	7.77%	15.00%	**55.65%**
N	6399	7044	6654	6100	7553

This table reports the transition frequency across carbon emissions from year 0 to year t in each panel. We first sort firms into five quintile groups based on their year 0 carbon emissions, and we report the probability that the firm should stay in this quintile group after 1/3/5/7 years. We bold the probability that the firm stays in the same group for each panel on the diagonal line. The columns indicate groups formed at year 0, and the rows in each panel indicate groups formed after 1/3/5/7 years. The last row in each panel indicates the number of observations within the quintile. The sample period is from 2002 to 2021.

Table 9: Carbon emission and inclusion into ESG-related fund

	Total inclusion			Average inclusion		
	(1)	(2)	(3)	(4)	(5)	(6)
LOGGHG	-0.0653*	-0.1326***	-0.1224***	-0.0001	-0.0036*	-0.0037*
	(-1.78)	(-2.98)	(-3.33)	(-0.26)	(-1.94)	(-1.87)
InstitOwn			0.0785***			0.0231***
			(-8.49)			(5.24)
Controls	T	T	T	T	T	T
Firm FE	T	T	T	T	T	T
Year FE		T	T		T	T
R2	0.50	0.52	0.57	0.68	0.71	0.72
N	67912	67912	67912	67912	67912	67912

In this table, we examine the relationship between carbon emission and the probability of being included in an ESG-related fund. An ESG-related fund is defined as funds with the keywords "CLEAN", "ESG", or "SOCIAL" in their fund names. The dependent variable is either the total number of being included in an ESG-related fund, or the probability of being included in an ESG-related fund, and the independent variables include the logarithmic value of firms' scope 1 carbon emissions, other firm fundamentals include firms' sales, total assets, non-current assets, firm size, leverage ratio, book-to-market ratio, investment ratio, ROE, HHI index, Plant, property & equipment, sales growth rate, and EPS growth rate. We also include institutional ownership in the regressions. We control for firms' year-fixed effects and industry-fixed effects in the regression. All standard errors are double clustered at the firm and year levels. The data period is from 2002 to 2021.

Table 10: Comparison of the determinants of carbon emission

	LOGGHG		GHGGR		GHGINTEN	
	(1)	(2)	(3)	(4)	(5)	(6)
LOGAT	0.5221***	0.2235***	-0.0058***	0.0170	-0.3374***	-21.5043***
	(19.34)	(5.33)	(-3.64)	(1.10)	(-5.22)	(-15.38)
B2M	0.0104	0.0258	0.0005	-0.0169	0.0024	2.7106***
	(1.22)	(1.56)	(0.32)	(-1.25)	(0.06)	(6.43)
ROE	0.3546***	-0.1557**	-0.0275***	-0.0484*	-0.2737**	-57.4743***
	(8.02)	(-2.04)	(-3.52)	(-1.67)	(-1.97)	(-12.28)
LEVERAGE	0.2730**	0.8724***	0.0094	-0.0151	0.7379**	-19.0196***
	(2.36)	(4.74)	(1.10)	(-0.54)	(2.46)	(-3.34)
INVEST2A	-1.6093***	-2.3727***	0.0176	-0.0518	-5.6307***	-130.9474***
	(-3.00)	(-3.02)	(0.16)	(-0.32)	(-2.68)	(-5.36)
HHI	0.7030	0.3381	0.3377***	-0.4911**	-0.5647	20.8282
	(1.47)	(0.39)	(3.69)	(-2.05)	(-0.70)	(0.65)
LOGPPE	0.3224***	-0.0252	0.0029	0.0006	0.3727***	-0.4947
	(11.73)	(-0.53)	(1.49)	(0.10)	(7.46)	(-0.44)
SALESGR	-0.0302	-0.0202	0.8888***	0.4510***	-0.1606	-16.9389***
	(-0.61)	(-0.26)	(19.88)	(5.10)	(-1.29)	(-5.18)
EPSGR	-0.0036	0.0109	-0.0025**	-0.0036	0.0178	2.4815***
	(-1.11)	(1.09)	(-2.01)	(-1.18)	(1.21)	(7.20)
Const	T	T	T	T	T	T
Year FE	T	T	T	T	T	T
Ind FE	T	T	T	T	T	T
R2	0.58	0.02	0.28	0.01	0.01	0.20
N	29146	67912	26089	60194	29143	67720
Data sample	Trucost	XGB	Trucost	XGB	Trucost	XGB

This table examines the determinants and compares carbon emissions estimated by the XGBoost algorithm and the original emission data provided by the Trucost database. The dependent variables use three different measures of carbon emission, including the logarithmic value of carbon emission, emission growth rate, and emission intensity defined as carbon emission over firm sales, on firm fundamentals. Independent variables include the logarithmic value of firms' total asset value, book-to-market ratio, leverage ratio, investment ratio, ROE, HHI index, Plant, property & equipment, sales growth rate, and EPS growth rate. In columns 1, 3, and 5 we use the original data sample provided by the Trucost database, and in the remaining columns, we use the data sample estimated by the XGBoost algorithm. We control both the year-fixed effect and industry-fixed effect. All standard errors are clustered at the industry level. The sample period is from 2002 to 2021.

Table 11: Main results: carbon emission and realized stock returns

	Trucost original sample						Carbon emissions estimated by XGBoost algorithm											
Sample period	Panel A: 2002-2016 (before the Paris agreement)						Panel B: 2002-2016 (before the Paris agreement)						Panel C: 2016-2021 (after the Paris agreement)					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
LOGGHG	0.0337	0.0620***					-0.0077	-0.0090					-0.0798**	-0.0606*				
	(1.44)	(3.06)					(-0.83)	(-0.92)					(-2.39)	(-1.81)				
GHGGR			0.6154***	0.6190***					0.0399**	0.0418**					0.0032	0.0414		
			(5.68)	(6.28)					(2.28)	(2.23)					(0.15)	(1.32)		
GHGINTEN					-0.0036	0.0010					-0.0029***	-0.0029***					-0.0066***	-0.0066***
					(-0.46)	(0.13)					(-4.80)	(-6.19)					(-5.04)	(-5.64)
LOGSIZE	-0.0936	-0.1115	-0.0437	-0.0250	-0.1028	-0.0877	0.0804	0.0655	0.1459	0.1358	0.0449	0.0228	0.3858**	0.3575**	0.4067**	0.3658**	0.3024	0.2359
	(-0.91)	(-1.02)	(-0.40)	(-0.20)	(-1.01)	(-0.78)	(0.83)	(0.70)	(1.51)	(1.35)	(0.46)	(0.25)	(2.01)	(2.08)	(2.10)	(2.11)	(1.63)	(1.51)
B2M	-0.0535*	-0.0490	-0.0413	-0.0207	-0.0569*	-0.0389	0.0242	0.0353	0.0286	0.0397	0.0193	0.0256	-0.0475	-0.0807	-0.0396	-0.0759	-0.0706	-0.1189
	(-1.81)	(-1.53)	(-1.37)	(-0.57)	(-1.92)	(-1.16)	(0.61)	(0.79)	(0.67)	(0.82)	(0.51)	(0.60)	(-0.60)	(-1.03)	(-0.51)	(-0.98)	(-0.87)	(-1.50)
LEVERAGE	0.0450	0.0276	0.0170	0.1488	0.0251	0.0893	0.0701	0.1790	0.1063	0.1996	0.0031	-0.0091	0.3866	-0.9289**	0.3933	-1.0218**	0.1925	-1.2516***
	(0.15)	(0.10)	(0.06)	(0.51)	(0.08)	(0.31)	(0.17)	(0.51)	(0.26)	(0.58)	(0.01)	(-0.03)	(0.70)	(-2.33)	(0.72)	(-2.18)	(0.37)	(-3.17)
MOM	0.0028	-0.0103	-0.0299	-0.0462	0.0037	-0.0103	-0.2236**	-0.2310**	-0.2195*	-0.2265*	-0.2272**	-0.2342**	-0.5265***	-0.5746***	-0.5186***	-0.5660***	-0.5246***	-0.5713***
	(0.02)	(-0.09)	(-0.24)	(-0.37)	(0.03)	(-0.09)	(-2.09)	(-2.13)	(-1.78)	(-1.82)	(-2.11)	(-2.15)	(-2.89)	(-3.09)	(-2.88)	(-3.09)	(-2.75)	(-2.96)
INVEST2A	-3.5287**	-2.7435***	-3.8919**	-3.2279***	-3.4364**	-3.0594***	-4.9874***	-4.6197***	-4.8443**	-4.3840***	-4.9874***	-4.7851***	-7.1192***	-5.3868*	-6.3466***	-4.1761	-8.2208***	-6.0479**
	(-2.14)	(-3.15)	(-2.09)	(-3.35)	(-2.05)	(-3.56)	(-2.74)	(-4.52)	(-2.59)	(-4.63)	(-2.73)	(-4.50)	(-3.07)	(-1.81)	(-2.95)	(-1.56)	(-3.54)	(-2.10)
ROE	0.6249***	0.5696***	0.6181**	0.5590**	0.6294***	0.5778***	1.6927***	1.6425***	1.5159***	1.4841***	1.4731***	1.4849***	3.0394***	2.4446***	3.0521***	2.4711***	2.7385***	2.2456***
	(2.68)	(2.77)	(2.61)	(2.61)	(2.64)	(2.81)	(7.53)	(7.75)	(6.87)	(6.93)	(7.02)	(7.19)	(7.67)	(13.68)	(7.34)	(12.14)	(6.76)	(12.16)
HHI	0.6989	0.9483	0.7801	1.0510	0.6511	0.9093	-0.6668	-2.1530	-0.7993*	-3.4369	-0.5328	-2.2067	-1.8179	-8.2538	-1.9053	-6.8391	-1.9716	-7.0282
	(1.04)	(0.78)	(1.26)	(0.60)	(0.98)	(0.75)	(-1.43)	(-0.99)	(-1.89)	(-1.42)	(-1.13)	(-1.03)	(-0.72)	(-0.48)	(-0.77)	(-0.44)	(-0.81)	(-0.42)
LOGPPE	-0.0194	-0.0231	0.0208	0.0107	0.0216	0.0079	0.1358**	0.1493***	0.0999	0.1109*	0.1045*	0.1252**	0.0243	0.1515	-0.0383	0.1162	-0.0395	0.1424
	(-0.43)	(-0.46)	(0.37)	(0.22)	(0.44)	(0.17)	(2.08)	(2.90)	(1.58)	(1.95)	(1.68)	(2.42)	(0.23)	(1.43)	(-0.31)	(1.01)	(-0.34)	(1.39)
BETA	-0.4926**	-0.4764**	-0.4551*	-0.4328*	-0.4978**	-0.4730**	-0.6403***	-0.6204***	-0.5839**	-0.5730**	-0.6634***	-0.6316***	-0.3516	-0.2189	-0.3338	-0.2104	-0.3520	-0.2177
	(-2.00)	(-1.98)	(-1.72)	(-1.68)	(-2.03)	(-1.97)	(-2.77)	(-2.78)	(-2.45)	(-2.50)	(-2.89)	(-2.83)	(-1.24)	(-0.73)	(-1.16)	(-0.68)	(-1.30)	(-0.75)
VOLAT	0.2023**	0.2126**	0.2043**	0.2167**	0.2015**	0.2118**	0.3086***	0.3137***	0.3070***	0.3127***	0.3128***	0.3181***	0.3964***	0.4239***	0.3925***	0.4197***	0.3949***	0.4221***
	(2.48)	(2.60)	(2.29)	(2.41)	(2.48)	(2.59)	(4.51)	(4.68)	(3.96)	(4.10)	(4.54)	(4.72)	(5.19)	(5.51)	(5.15)	(5.47)	(5.03)	(5.35)
SALESGR	-0.2617	-0.2837	-0.4227	-0.4292	-0.2583	-0.2842	-0.4954***	-0.4646***	-0.5312***	-0.4908***	-0.5487***	-0.5452***	-0.7597**	-0.6410*	-0.7678*	-0.6818**	-0.8133**	-0.7311**
	(-1.00)	(-1.10)	(-1.50)	(-1.52)	(-0.99)	(-1.10)	(-3.28)	(-3.03)	(-3.12)	(-2.80)	(-3.90)	(-3.80)	(-2.04)	(-1.96)	(-1.94)	(-1.97)	(-2.14)	(-2.22)
EPSGR	0.0207	0.0252	0.0033	0.0084	0.0207	0.0252	0.0950***	0.0975***	0.0770***	0.0782***	0.1064***	0.1062***	0.0309	0.0356	0.0266	0.0319	0.0420	0.0417
	(0.75)	(0.94)	(0.15)	(0.39)	(0.74)	(0.94)	(3.77)	(3.82)	(2.95)	(2.97)	(3.94)	(3.92)	(0.70)	(0.88)	(0.58)	(0.77)	(1.01)	(1.10)
Const	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T
Year-Mon FE	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T
Ind FE		T		T		T		T		T		T		T		T		T
R2	0.25	0.25	0.25	0.26	0.25	0.25	0.16	0.16	0.17	0.17	0.16	0.16	0.16	0.17	0.17	0.17	0.17	0.17
N	172059	172059	157693	157693	172011	172011	533001	533001	457178	457178	531136	531136	231149	231149	223551	223551	229777	229777

This table examines the relation between carbon emission and stock realized returns with different samples. The dependent variable is the stock return of firm i in year t. The variables of interest are three measures of firms' carbon emissions, including the logarithmic value of carbon emission, emission growth rate, and emission intensity defined as carbon emission over firm sales. Other firm fundamentals include firm size, book-to-market ratio, leverage ratio, momentum, investment ratio, ROE, HHI index, Plant, property & equipment, Beta, return volatility, sales growth rate, and EPS growth rate. We use three different samples for empirical estimation. The first sample is the dataset obtained from the Trucost database, with a sample period from 2002 to 2016, and has 215808 observations. The second sample is emission data predicted from the XGBoost model with a partial sample period from 2002 to 2016. It has 533001 observations. The third sample period focuses on a later XGBoost estimated sample period from the start of 2016 to 2021. We control for year-month fixed effects and include industry-fixed effects separately in the regressions, and we cluster standard errors at the 2-digit GIC industry and year levels.

Table 12: Carbon risk premium by year

Panel A: Estimated by LOGGHG											
IndepVar.	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	Avg coef
LOGGHG	-0.0551*	0.0515*	-0.0309	-0.0024	-0.0343**	-0.0109	-0.0441*	0.0784***	0.0069	0.0216	-0.0019
	(-1.70)	(1.90)	(-1.44)	(-0.15)	(-2.17)	(-0.67)	(-1.76)	(2.66)	(0.35)	(1.22)	(-0.15)
R2	0.02	0.07	0.03	0.02	0.02	0.02	0.02	0.12	0.01	0.02	
N	25186	32001	35348	42477	42026	41312	41096	40719	40637	39633	
	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	Avg coef
LOGGHG	0.0181	0.0061	0.0037	-0.0241	-0.0685***	-0.0588***	-0.0685***	-0.0866***	-0.2139***	0.0619**	-0.0431***
	(1.02)	(0.36)	(0.22)	(-1.30)	(-2.84)	(-2.81)	(-3.02)	(-3.30)	(-5.42)	(2.11)	(-1.88)
R2	0.02	0.03	0.02	0.01	0.03	0.01	0.01	0.03	0.12	0.10	
N	38634	38258	38141	37533	38264	38728	38480	38225	38633	38819	
Panel B: Estimated by GHGGR											
IndepVar.	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	Avg coef
GHGGR		0.0658	0.0724	-0.0402	0.0454	0.0192	0.0560	-0.0067	0.0910	-0.0241	0.0310**
		(0.76)	(1.22)	(-0.78)	(0.91)	(0.34)	(0.63)	(-0.06)	(1.38)	(-0.37)	(2.13)
R2		0.07	0.02	0.02	0.02	0.02	0.02	0.12	0.01	0.02	
N		23745	29697	36027	38000	37984	37984	37581	37723	36836	
	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	Avg coef
GHGGR	0.0149	0.0050	0.0936	0.0565	-0.2562**	-0.0211	0.0179	-0.0278	0.2148*	0.0551	0.0153
	(0.21)	(0.08)	(1.62)	(0.80)	(-3.28)	(-0.32)	(0.27)	(-0.40)	(1.92)	(1.60)	(0.429)
R2	0.02	0.03	0.01	0.01	0.03	0.01	0.01	0.03	0.12	0.10	
N	35670	35328	35462	35141	34278	37872	37691	37522	37872	38316	
Panel C: Estimated by GHGINTEN											
IndepVar.	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	Avg coef
GHGINTEN	-0.0067***	0.0007	-0.0012*	-0.0035***	-0.0028***	-0.0014***	-0.0010	-0.0029***	-0.0032***	-0.0023***	-0.0024***
	(-5.89)	(0.87)	(-1.84)	(-6.65)	(-5.63)	(-2.82)	(-1.21)	(-2.97)	(-5.02)	(-3.94)	(-2.54)
R2	0.02	0.07	0.03	0.02	0.02	0.02	0.02	0.12	0.01	0.02	
N	25117	31893	35240	42405	41930	41171	40964	40599	40501	39465	
	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	Avg coef
GHGINTEN	-0.0024***	-0.0025***	-0.0029***	-0.0004	-0.0071***	-0.0022***	-0.0021***	-0.0068***	-0.0090***	-0.0054***	-0.0041***
	(-4.13)	(-4.67)	(-5.14)	(-0.60)	(-8.82)	(-3.12)	(-2.79)	(-8.17)	(-6.46)	(-5.06)	(-4.84)
R2	0.02	0.03	0.02	0.01	0.03	0.01	0.01	0.03	0.12	0.10	
N	38421	38080	37973	37377	38180	38477	38300	37992	38333	38495	

This table examines the time-varying relationship between carbon emission and stock returns. We estimate cross-sectional regressions of stock returns on firms' carbon emissions following 5. The dependent variable is the stock return of firm i in year-month t, and the independent variables of interest are three measures of firms' carbon emissions, including the logarithmic value of carbon emission, emission growth rate, and emission intensity defined as carbon emission over firm sales, which are reported in panel A, B, and C, respectively. We control for financial variables, including firm size, book-to-market ratio, leverage ratio, momentum, investment ratio, ROE, HHI index, Plant, property & equipment, Beta, return volatility, sales growth rate, and EPS growth rate. We also control for industry-fixed effects and cluster standard errors at the 2-digit GIC industry and year levels. We report estimated coefficients of carbon emission in each year's regression. In the column to the right, we report the average premium during the sub-sample period with t-statistics. The sample period is from 2002 to 2021.

Table 13: Mechanism: Carbon emission and investor flow

Dep Var.	LOGFLOW					
	(1)	(2)	(3)	(4)	(5)	(6)
POST*GHG	-0.0494***	-0.0492***	-0.0267*	-0.0383***	0.0001	0.0000
	(-5.55)	(-4.96)	(-1.92)	(-3.03)	(0.17)	(0.15)
LOGGHG	0.0082*	0.0058				
	(1.82)	(1.33)				
GHGGR			0.0354***	0.0361***		
			(5.90)	(6.04)		
GHGINTEN					-0.0019***	-0.0017***
					(-11.25)	(-11.02)
POST	0.2416	0.2303	-0.2601*	-0.2620*	-0.2925**	-0.2984**
	(1.58)	(1.40)	(-1.82)	(-1.84)	(-2.06)	(-2.11)
LOGSIZE	1.1038***	1.0854***	1.0905***	1.0654***	1.0825***	1.0626***
	(36.15)	(30.91)	(36.80)	(30.59)	(34.93)	(29.89)
B2M	0.0918***	0.1046***	0.0896***	0.0998***	0.0879***	0.0984***
	(6.58)	(7.01)	(6.72)	(6.80)	(6.48)	(6.54)
LEVERAGE	-0.5430***	0.1918*	-0.5245***	0.1652	-0.5405***	0.1143
	(-5.08)	(1.79)	(-4.93)	(1.57)	(-5.01)	(1.04)
INVEST2A	1.8859***	1.7273***	1.8980***	1.7840***	1.8322***	1.7185***
	(3.61)	(3.42)	(3.18)	(3.22)	(3.40)	(3.41)
ROE	-0.5043***	-0.3049***	-0.4893***	-0.2891***	-0.6095***	-0.3861***
	(-9.81)	(-6.95)	(-9.86)	(-6.61)	(-11.45)	(-8.56)
HHI	0.1984	0.2923	0.2119	0.1115	0.2637	0.2200
	(0.55)	(0.53)	(0.55)	(0.22)	(0.73)	(0.41)
LOGPPE	-0.0072	-0.0059	-0.0175	-0.0077	-0.0266	-0.0188
	(-0.36)	(-0.22)	(-0.93)	(-0.28)	(-1.42)	(-0.70)
SALESGR	0.2934***	0.2602***	0.2975***	0.2702***	0.2967***	0.2528***
	(3.74)	(3.62)	(3.31)	(3.32)	(3.87)	(3.53)
EPSGR	-0.0358***	-0.0387***	-0.0376***	-0.0405***	-0.0307**	-0.0352***
	(-2.65)	(-2.90)	(-3.07)	(-3.34)	(-2.45)	(-2.77)
Const	T	T	T	T	T	T
Ind FE		T		T		T
R2	0.67	0.66	0.67	0.65	0.68	0.66
N	29602	29602	27084	27084	29532	29532

This table examines the relationship between firms' carbon emissions and institutional investor flow before and after the ratification of the Paris Agreement. The dependent variable is the firms' logarithmic value of investor flow, and the independent variable is a cross term that interacts with firms' carbon emissions and a time dummy that denotes Paris Agreement. We include three different measures of carbon emissions including the logarithmic value of carbon emission, emission growth rate, and emission intensity defined as carbon emission over firm sales. We control for financial variables, including firm size, book-to-market ratio, leverage ratio, momentum, investment ratio, ROE, HHI index, Plant, property & equipment, Beta, return volatility, sales growth rate, and EPS growth rate. We also control for industry-fixed effects and cluster standard errors at the 2-digit GIC industry and year levels. We report estimated coefficients of carbon emission in each year's regression. In the column to the right, we report the average premium during the sub-sample period with t-statistics. The sample period is from 2002 to 2021.

Table 14: Flow-induced stock returns

Dep Var.	RET									
Portfolio	Lo		2		3		4		Hi	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A: Sorted by LOGGHG										
LOGFLOW	1.4940***	1.4321***	0.4999	0.5158	0.8970	0.8713	1.0346*	1.0432*	1.6189	1.6789
	(4.31)	(4.33)	(1.20)	(1.35)	(1.13)	(1.04)	(1.93)	(1.85)	(1.18)	(1.21)
Const	T	T	T	T	T	T	T	T	T	T
Controls	T	T	T	T	T	T	T	T	T	T
Year-Mon FE	T	T	T	T	T	T	T	T	T	T
Ind FE		T		T		T		T		T
R2	0.18	0.19	0.22	0.23	0.24	0.25	0.22	0.22	0.27	0.28
N	15660	15660	15656	15656	15657	15657	15656	15656	15659	15659
Panel B: Sorted by GHGGR										
LOGFLOW	1.4931***	1.4299***	0.4921	0.5095	0.9062	0.8816	1.0373*	1.0460*	1.6158	1.6805
	(4.30)	(4.32)	(1.17)	(1.32)	(1.15)	(1.06)	(1.94)	(1.85)	(1.18)	(1.21)
Const	T	T	T	T	T	T	T	T	T	T
Controls	T	T	T	T	T	T	T	T	T	T
Year-Mon FE	T	T	T	T	T	T	T	T	T	T
Ind FE		T		T		T		T		T
R2	0.18	0.19	0.22	0.23	0.24	0.25	0.21	0.22	0.27	0.28
N	15660	15660	15656	15656	15657	15657	15656	15656	15659	15659
Panel C: Sorted by GHGINTEN										
LOGFLOW	1.4846***	1.4223***	0.5146	0.5308	0.9038	0.8744	1.0312*	1.0392*	1.6202	1.6819
	(4.27)	(4.28)	(1.26)	(1.42)	(1.13)	(1.04)	(1.93)	(1.85)	(1.18)	(1.21)
Const	T	T	T	T	T	T	T	T	T	T
Controls	T	T	T	T	T	T	T	T	T	T
Year-Mon FE	T	T	T	T	T	T	T	T	T	T
Ind FE		T		T		T		T		T
R2	0.18	0.19	0.22	0.23	0.24	0.25	0.22	0.22	0.27	0.28
N	15660	15660	15656	15656	15657	15657	15656	15656	15659	15659

This table examines the relation between institutional investor flow and stock realized returns for stocks with different emission levels. We first sort stocks of firms based on their logarithmic carbon emissions into 5 quintiles and test the heterogeneity of flow-induced stock returns. The sorting variables are three different measures of carbon emissions including the logarithmic value of carbon emission, emission growth rate, and emission intensity defined as carbon emission over firm sales in panels A, B, and C, respectively. The dependent variable is stock returns and the independent variables are the logarithmic value of investor flows. Other firm fundamentals include firm size, book-to-market ratio, leverage ratio, momentum, investment ratio, ROE, HHI index, Plant, property & equipment, Beta, return volatility, sales growth rate, and EPS growth rate. We restrict the sample period to the date after the ratification of the Paris Agreement. We control for year-month fixed effects and include industry-fixed effects separately in the regressions, and we cluster standard errors at the 2-digit GIC industry and year levels.

Table 15: The carbon premium and common risk factors

	Panel A: 2002-2021						Panel B: 2016-2021					
	LOGGHG		GHGGR		GHGINTEN		LOGGHG		GHGGR		GHGINTEN	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Intercept	-0.0301**	-0.0197*	0.0393**	0.0262	-0.0027***	-0.0028***	-0.0853***	-0.0700***	0.0335	0.0077	-0.0045***	-0.0045***
	(-2.41)	(-1.67)	(2.17)	(1.53)	(-5.03)	(-6.68)	(-5.16)	(-2.97)	(0.71)	(0.16)	(-4.30)	(-5.61)
RMRF	0.0019		-0.0046		-0.0001		0.0054*		-0.0135		-0.0001	
	(0.94)		(-0.99)		(-1.38)		(1.85)		(-1.42)		(-0.56)	
SMB	-0.0129***		-0.0014		-0.0007***		-0.0284***		-0.0065		-0.0012***	
	(-2.82)		(-0.13)		(-4.47)		(-4.91)		(-0.24)		(-3.72)	
HML	-0.0017		-0.0141		-0.0003		0.0061		-0.0269**		-0.0001	
	(-0.27)		(-1.43)		(-1.44)		(0.63)		(-1.99)		(-0.28)	
RMW	0.0039		-0.0172		-0.0001		0.0096		-0.0160		0.0000	
	(0.94)		(-1.40)		(-0.41)		(1.41)		(-0.70)		(0.01)	
CMA	0.0045		-0.0183*		-0.0007***		0.0027		-0.0189		-0.0006**	
	(0.63)		(-1.87)		(-3.17)		(0.17)		(-1.15)		(-2.01)	
BAB	0.0124***		-0.0060		0.0003**		0.0159**		-0.0230*		0.0003	
	(3.96)		(-0.85)		(2.33)		(2.40)		(-1.95)		(0.89)	
LIQ	-0.0001		-0.0053		-0.0002**		0.0016		0.0098		-0.0000	
	(-0.02)		(-1.08)		(-1.98)		(0.36)		(0.85)		(-0.28)	
Mom	-0.0014		0.0022		-0.0001		-0.0006		0.0055		-0.0002	
	(-0.67)		(0.52)		(-0.76)		(-0.17)		(0.49)		(-0.82)	
R2	0.19	0.00	0.07	0.00	0.26	0.00	0.32	0.00	0.24	0.00	0.24	0.00
N	240	240	228	228	240	240	72	72	72	72	72	72

This table examines the carbon return premium after controlling for common risk factors. The dependent variable is the monthly carbon premium estimated using a cross-sectional return regression in 5. We use three measures of firms' carbon emissions, including the logarithmic value of carbon emission, emission growth rate, and emission intensity defined as carbon emission over firm sales. We apply different adjustments for risk exposure by performing the Fama-Macbeth regression of monthly risk premia on common risk factors, including the market factor as the CAPM model in panel A with a full sample estimated with the XGBoost data. In panel B, we restrict the time period to 2016 to 2021. Other factors from other widely adopted models like the Fama-French three-factor and the five-factor model, the three-factor model with Carhart's Momentum Factor Carhart (1997), and Pastor-Stambaugh's liquidity factor Pástor and Stambaugh (2003), a Betting-Against-Beta factor in Frazzini and Pedersen (2014). We control for 2-digit GIC industry-fixed effects. We also calculate the standard errors of the coefficients using the Newey-West robust estimator with 12 lags to adjust serial correlations.

Appendix A. Robustness analyses

A.1. Business structures and carbon emission patterns

Despite the intuitiveness, it remains to be investigated whether firms with similar business structures share similar emission patterns. There may exist industry-specific heterogeneity that prevents the algorithm from making accurate carbon emission predictions. For example, some firms might rely on cutting-edge technologies to manufacture renewable and clean products, whereas their rival firms are still using heavy-pollution technologies for manufacturing. Admittedly, firms that have the incentive to reduce carbon emissions are more competitive. On the other hand, these firms are more likely to disclose carbon emissions and have better financial performance as well as growth opportunities. On the contrary, non-disclosure firms would produce more carbon emissions. This could bias our estimation of carbon emission downward.

We designed two tests to examine whether business similarity helps to predict carbon emissions. In the first test, we only include firms' business similarity scores and their fixed effects (GVKEY) in both the training set and the test set. We use the same training method as well as machine learning parameters as in the main analyses and rerun the test, and we remove all the other firm fundamentals including sales, total assets, non-current assets, and PPENT for both firms from the training set. As subfigure A in figure A1 shows, the out-of-sample R-square declined a little bit from 0.81 in the main analysis (see subfigure B in figure 3) to 0.68 with a marginal declination of 0.13. This result shows that only including the similarity score in the algorithm has strong predictive power for carbon emissions.

We also assign random business similarity score pairs to the original pairs in subfigure B in figure A1. We keep other settings unchanged and retrain the algorithm. Results suggest that the inclusion of random business similarity scores lowers the R-square to 0.64, with a marginal declination of 0.17. We do not observe a very low R-square because we include both the firm fixed effects and firm fundamentals in the algorithm.

[Insert Figure A1 near here]

A.2. Different sample partitioning method

In our main analysis, we partition data into a training set and a test set by year. We set observations from 2002 to 2018 as the training set, and we set observations from 2019 to 2021 as the test set. We train the XGBoost algorithm based on this set of data and predict GHG subsequently. However, this partitioning method is not random and it is subject to severe sampling bias, as carbon emissions might be affected by unobserved time trends. As a result, we partition the sample period either from 2002 to 2017 as the

training set and 2018 to 2021 as the test set or by 2002 to 2019 as the training set and 2020 to 2021 as the test set.

We perform similar training and validating approaches with these two different training and testing approaches. We first report results with a training set spanning from 2002 to 2017 and a test set spanning from 2017 to 2021 in subfigures (a) and (b) in A2. We plot the learning curve for both the training set and the test set in subfigure (a), where two curves become flattened after two thousand times of iterations and remain steady afterward. In subfigure (b), we report the out-of-sample prediction vs. real emission for disclosed firms. The average R2 is 0.79 which suggests that the algorithm makes a good fit for the carbon emissions disclosed by US firms, which suggests that our partition method is reliable.

In figure A3, we plot variable importance and contribution with an important plot and a SHAP value plot. In subfigure (a) and (b) where the training set is from 2002 to 2017, we plot the importance plot, where the sequence of important variables remains unchanged, as the two most important variables are firms' business similarity scores and the carbon emission of disclosed firms. Moreover, firm fundamentals for non-disclosure firms are more important than firm fundamentals for disclosure firms. In panel B, we plot the SHAP value plot.

[Insert Figure A2 and A3 near here]

Then, we partition the period by setting observations from 2002 to 2018 as the training set, and the year 2019 as the test set. We perform similar tests in the remaining subfigures in figure A2 and A3, and training results are largely the same.

A.3. Cross-validation test

In this section, we use traditional machine-learning methods cross-validated our model parameters. We report the results when we set different learning rates and tree depths of the model. We set the evaluation metric as the R2 of the in-sample training results. In figure A4, the result shows that the learning rate is optimized over 0.2 and tree depth optimized over 7, which are all parameters we have set when training the original XGBoost model. We also cross-validated other model parameters, and in-sample R2 is strikingly high.

[Insert Figure A4 near here]

A.4. Comparison with linear models

We also compare XGBoost-trained results with linear models. In Rossi and Utkus (2021), the authors compare the performance between Boosted regression trees models

and linear regression models with a cross-validation analysis. Following their approach, we bootstrap 75% of the training set and test set with 300 times of iterations. The original training set is the one we used in our main analysis, with a sample period spanning from 2002 to 2018. We do so to assess whether there should be overfitting problems in the non-parametric model. Our model parameters are identical to the default settings, with the evaluation objective as RMSE, and we set other input variables like sale, non-current asset, total asset, and employees as the same. We report the R2 of the model as another metric. Sampling results give distributions of R2 for out-of-sample models, which we report in sub-figure A of figure A5. As can be seen from the figure, for out-of-sample R2, the distribution plot is centered around the mean value of 0.77, which is slightly lower than that of the baseline results.

We also made a comparison between our XGBoost-based machine learning model and traditional linear models to show the strength of non-parametric estimation in this case. Although linear models are easier to interpret and be explained, machine learning models tend to outperform on average. We follow a similar methodology by bootstrapping 300 times our original training set. We train the linear model which has identical covariates as the XGBoost models, and we compute R2 for out-of-sample models at each iteration. We plot the density plot in sub-figure B of figure A5.

As can be seen from this sub-figure, the out-of-sample R2 for the linear model is around 0.19, which is significantly lower than that of the XGBoost. The comparison between the two plots suggests that in our case machine learning model does outperform traditional linear models, and they make better estimations both in the sample and out of the sample. We do not perform prediction comparisons with other regression or machine learning models like Neural networks or Random forests. We leave that for future research.

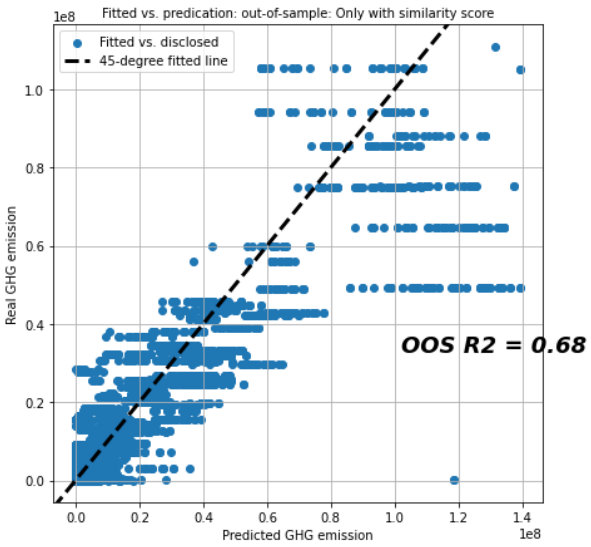
[Insert Figure A5 near here]

A.5. Emission autocorrelation patterns

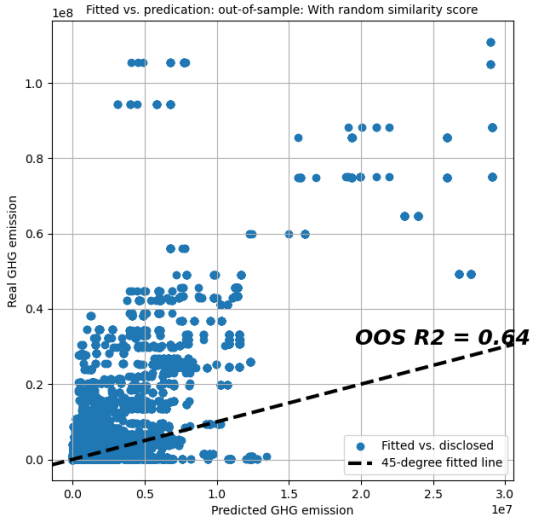
We show that carbon emission is quite persistent with data sample estimated by the XGBoost algorithm in table 8, as firms are more likely to stay within the emission quintiles 1/3/5/7 years after the formation date. We also perform auto-correlation regressions to examine the persistence of carbon emissions. We regress three different measures of carbon emissions, including the logarithmic value of carbon emission, emission growth rate, and emission intensity defined as carbon emission over firm sales, on their lagged variables, controlling for firm fundamentals and year-fixed effects. We double cluster standard errors at firm and year levels. In table A1, we report auto-regression results. Regression results suggest that the persistent relationship is quite significant. For log-

arithmic carbon emission data, the regression coefficient is 0.7585 and 0.7409 without and with controls, and t-statistics are 33.07 and 28.43, respectively. Emission intensity is persistent as well in columns 5 and 6. However, as shown in the main analyses, the emission growth rate is not persistent, as we are doing estimation on cross-sectional levels and do not correct business similarities on a time-series level.

[Insert Table A1 near here]

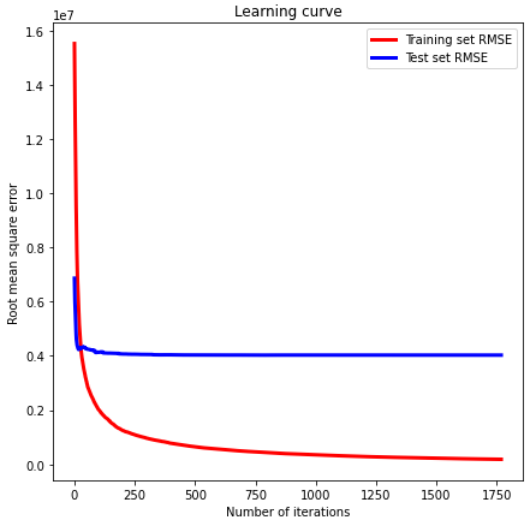


Subfigure A: Prediction only with similarity score

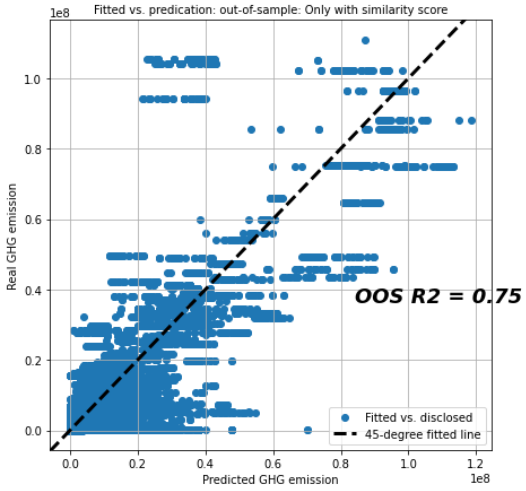


Subfigure B: Prediction with random similarity score

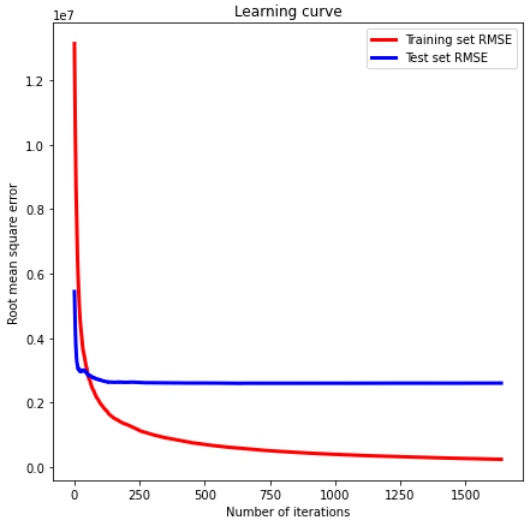
Fig. A1. Business similarity and emission similarity. In subfigure A, we only include firms' business similarity scores and their fixed effects (GVKEY) in both the training set and the test set. We use the same training method as well as machine learning parameters as in the main analyses and rerun the test, and we remove all the other firm fundamentals including sales, total assets, non-current assets, and PPENT for both firms from the training set. In subfigure B, we assign random business similarity score pairs to the original pairs and rerun the algorithm.



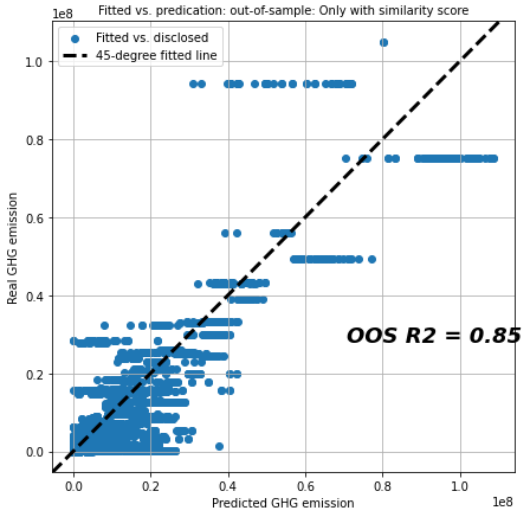
(a) Learning curve (Training set: 2002-2017)



(b) OOS validation (Training set: 2002-2017)

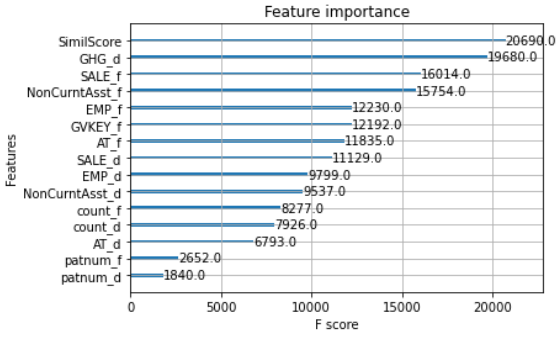


(c) Learning curve (Training set: 2002-2019)

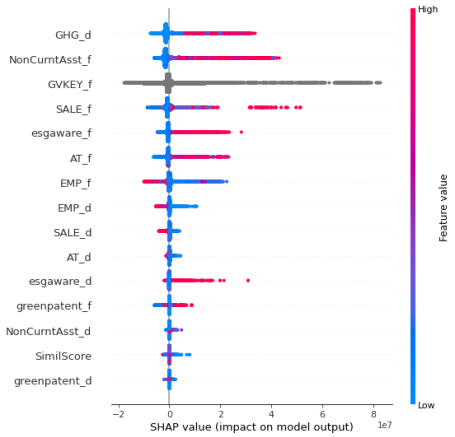


(d) OOS validation (Training set: 2002-2019)

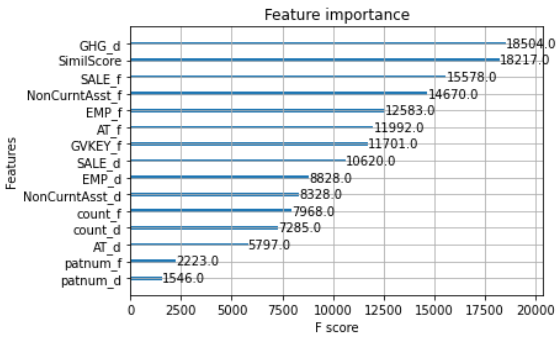
Fig. A2. XGBoost performance results with different training periods. In subfigures (a) and (b), we set the training set from 2002 to 2017, and the test set from period 2018 to 2021. In subfigures (c) and (d), we choose the training set from 2002 to 2019, and the test set from the period 2020 to 2021. In subfigures (a) and (c), we report the Root Mean Squared Error curve for both the training set and the validating set for the XGBoost model after 2 thousand times iterations, where the blue line denotes the test curve and the red line denotes the training curve. In subfigures (b) and (d), we report the out-of-sample validation tests. We use models trained from in-sample data to predict out-of-sample carbon emissions, and we compare the predicted out-of-sample values with real out-of-sample values. The x-axis indicates predicted GHG (scope 1 greenhouse gas emission), and the y-axis indicates real carbon emissions that are disclosed by firms or computed by the Trucost database. We add a 45-degree line to illustrate the fitness of our model.



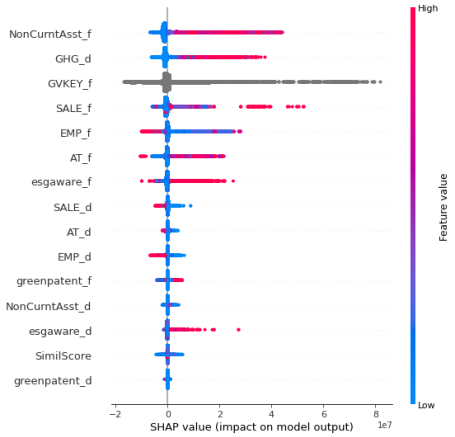
(a) Importance plot (Training set: 2002-2017)



(b) SHAP plot (Training set: 2002-2017)

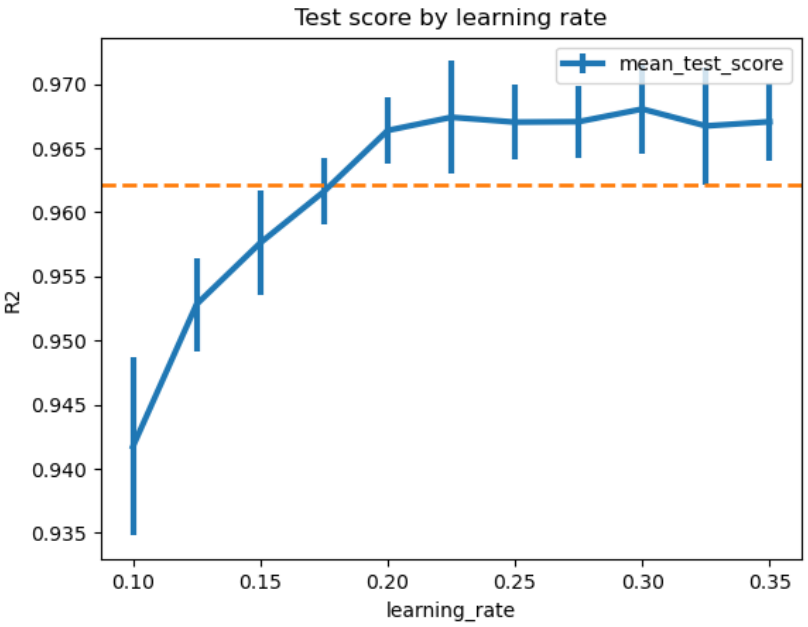


(c) Importance plot (Training set: 2002-2019)

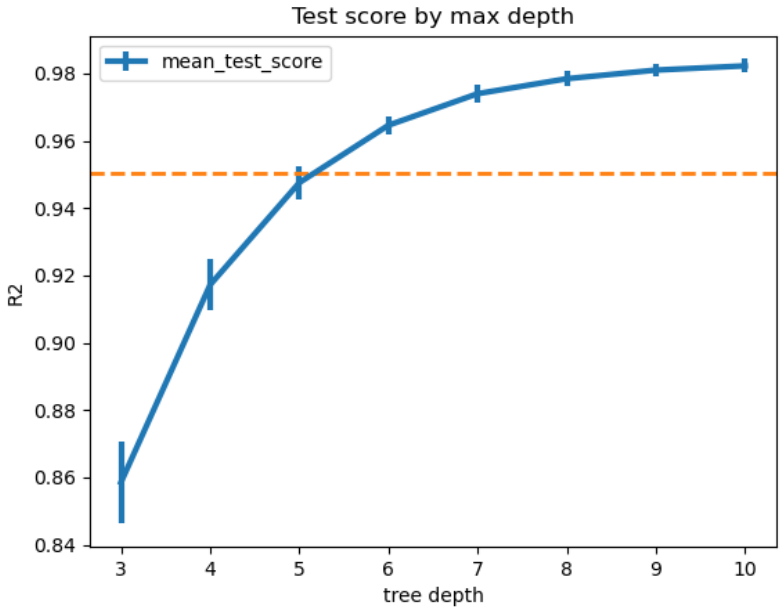


(d) SHAP plot (Training set: 2002-2019)

Fig. A3. Variable importance contribution plot with different training periods. In subfigures (a) and (b), we set the training set from 2002 to 2017, and the test set from period 2018 to 2021. In subfigures (c) and (d), we choose the training set from 2002 to 2019, and the test set from the period 2020 to 2021. We plot both the importance plot as well as the SHAP value plot for variables trained in the XGBoost model. In subfigures (a) and (c), we illustrate the importance of each variable identified by the machine learning algorithm. The importance is measured by each feature's percentage of total predictive power on the x-axis. The name of each feature is on the y-axis, where the most important four features are similarity scores between two firms, the carbon emission of the disclosed firm, non-current assets for the target firm, and sales for the non-disclosure firm. The higher the feature importance, the stronger predictive power the variable has. In subfigures (b) and (d), We present the SHAP value of each variable in the XGBoost model, which is a unified approach to explain the output in most tree models. The values in the x-axis show predictive power with positive or negative directions. Each dot represents an observation within the model. Higher inputs tend to have a higher SHAP value; a higher SHAP value means more importance or contribution to the model. All variables are displayed sequentially by their importance from top to bottom.

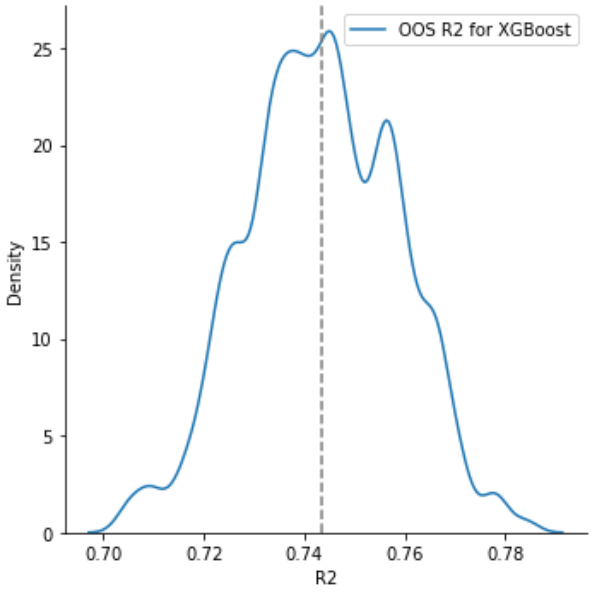


Subfigure A: Cross-validation test on learning rate

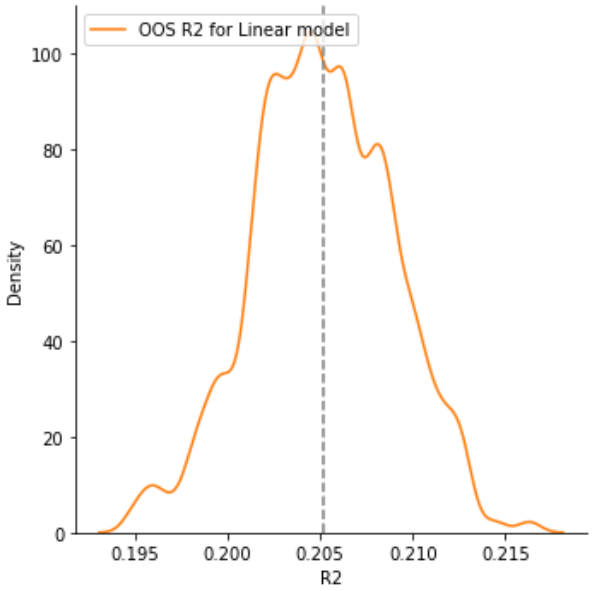


Subfigure B: Cross-validation test on tree depth

Fig. A4. Cross-validation by learning rate and tree depth. We perform additional tests on our XGBoost models by applying cross-validation with different parameters for learning rate and tree depth. We set five holds for in-sample training and report mean scores (with standard deviation) under different parameters for each in-sample training result. The evaluation metric is R2 for each panel. The yellow horizontal line reports the average test scores across different groups.



Panel A: XGBoost model



Panel B: Linear model

Fig. A5. Model comparison: XGBoost versus linear models. We randomly select 75% of the original training set and test set with 300 iterations to train the XGBoost model in subfigure A. We use R2 as the evaluation metric. To compare the XGBoost model with traditional linear models, which is reported in subfigure B, we use the same variables as covariates in an OLS model, where the dependent variable is carbon emission for the target firms, and the independent variables include cosine similarity scores, the carbon emission of the disclosed firm, and other firm fundamentals from both firms. The x-axis denotes the value of R2, and the y-axis denotes a probability density function.

Table A1: Emission persistency with auto-correlation test

	LOGGHG		GHGGR		GHGINTEN	
	(1)	(2)	(3)	(4)	(5)	(6)
LOGGHG	0.7585***	0.7409***				
	(33.07)	(28.43)				
GHGGR			-0.0982***	-0.1082***		
			(-11.87)	(-14.94)		
GHGINTEN					0.8359***	0.7991***
					(54.52)	(47.48)
Const	T	T	T	T	T	T
Control		T		T		T
Year FE	T	T	T	T	T	T
R2	0.58	0.57	0.01	0.02	0.74	0.73
N	76113	63463	62272	54582	71256	63283

This table examines the persistence with three measures of carbon emissions, including the logarithmic value of carbon emission, emission growth rate, and emission intensity defined as carbon emission over firm sales. We regress each measure on its lagged variables, controlling for firm fundamentals, including firm size, book-to-market ratio, leverage ratio, investment ratio, ROE, HHI index, Plant, property & equipment, sales growth rate, and EPS growth rate. We also add year-fixed effects in the regressions. All standard errors are clustered at both firm and year levels. The sample period is from 2002 to 2021.

Appendix B. Supplementary asset pricing results

B.1. Emission-return relationship with different sample periods

Since the Trucost data source does not include a large quantity of emission data prior to 2016, and many of the firms that exist in the database tend to be larger, more profitable, and produce more carbon emissions (see table 4), we do not regress stock returns on carbon emission with this unbalanced data set in our main empirical analysis. In other words, the positive emission-return relationship is more pronounced for the sample period prior to the Paris Agreement, which may overshadow the recent awareness of sustainable investment in recent years, and the regression results just may be unreliable and unconvincing.

As a result, in table B1, we estimate the emission-return relationship with either the data sample provided by Trucost or the one estimated by the XGBoost algorithm from 2002 to 2021. There is a total of 332410 observations for firm-month observations with Trucost data and a total of 764150 observations for the XGBoost estimated data.

[Insert Table B1 near here]

Regression results show that, in Panel A, where we use the whole sample period with Trucost data, the baseline regression result in column 1 is insignificant with coefficients of 0.0098 and t-statistics of 0.48. The positive relationship is more pronounced as we control for industry-fixed effects in column 2. The estimated coefficient becomes 0.0536 and the t-statistic of 2.34. This may be a result of the high with-in-industry emission-return relationship prior to 2016. In panel B where we use the XGBoost estimated data, the (significantly) positive relationship dissipates swiftly in columns 7 and 8, with coefficients -0.0239 and -0.0226 and t-statistics of -1.56 and -1.41, respectively.

In columns 3 to 4 and 9 to 10 where the independent variable is the emission growth rate, the positive relationship remains relatively stable for the Trucost sample but not the XGBoost estimated sample. As we show in the next section with sorting results (in table B3), higher emission growth rates are largely associated with higher sales growth rates, which largely is related to firms' growth prospects. In the remaining columns where we switch the emission intensity ratios, regression results are again negative but insignificant for the Trucost sample but consistently significant for the intensity measures. In unreported regression results where we normalize carbon emission with other firm fundamentals in the Trucost data sample, there also exhibits a negative but insignificant relationship. The inconsistency in regression results purports our earlier claim that it is delicate to choose proxies for firms' carbon emission risk.

We show that, in table B2, the Trucost sample also suggests a negative emission premium post the Paris Agreement, but the effect is not as pronounced as the ones

estimated with the XGBoost algorithm. In column 2, the regression coefficient shrinks from -0.0611 in column 1 to 0.0184. This marginal decrease both in terms of economic and statistical significance shows the comparative advantage of the emission data estimated by XGBoost.

[Insert Table B2 near here]

Overall, the main message we want to convey is that the negative carbon premium is much more significant post the Paris Agreement. The previously documented positive emission-return relationship is mostly notable prior to 2016 or even 2012. As more firms choose to disclose their carbon emissions and the Trucost database accumulates more data, we expect to observe different results in the coming years.

B.2. Portfolio sorting results

We examine cumulative portfolio returns of different quintiles sorted by logarithmic carbon emissions from 2002 to 2021 with data samples estimated with XGBoost or provided by the Trucost database. We report cumulative returns for value-weighted or equal-weighted hi-lo portfolios over the sample period in table B3. Sorting results suggest that on average high-emission stocks underperform low-emission stocks by 0.2248% and 0.1889% per month for value-weighted or equal-weighted portfolios, respectively. However, the negative premium is not significant as we are using the whole data period. We also report carbon emissions and firm fundamentals, including sales, total assets, non-current assets, firm size, leverage ratio, book-to-market ratio, investment ratio, ROE, HHI index, Plant, property & equipment, sales growth rate, and EPS growth rate. In contrast to previous works, this table shows that low-emission firms tend to be less profitable and have higher sales growth rates. This suggests that, apart from the firm size, carbon emission is highly correlated with firms' profit margins and growth prospects. In panel B, we report sorting results with data samples obtained from the Trucost database. As compared to the results in panel A, the hi-minus-low portfolio returns are significantly negative in the bottom line. This may be because the Trucost database provides more estimation post-2016, which makes the low-carbon premium more significant.

[Insert Table B3 near here]

B.3. Carbon emission and non-ESG risks

One potential and alternate mechanism behind this low-carbon premium phenomenon after 2016 may be driven by other sources of risks. One may concern that, after the Paris Agreement, there might arise other related market-level or firm-level risks, which are irrelevant to the carbon risk. On the other hand, adopting carbon-reducing technology might

introduce other sources of risks, as new manufacturing methods often come along with massive uncertainty and high costs. This naturally makes firms that adopt low-carbon styles riskier than other firms, and as a result, investors require higher compensation for holding these low-carbon firms. In that sense, we are not observing a transient positive shock on people's environmental beliefs, but rather an ordinary risk story.

To examine this hypothesis, we examine the relationship between carbon emission and other firms' fundamentally related risks post the Paris Agreement. We regress various financial ratios on firms' carbon emissions, and we interact the carbon emission with a dummy that indicates the date after the Paris Agreement. The variable of interest is the coefficient in front of the interaction term. If carbon emission is positively related to various sources of risks, then we cannot reject the alternative risk-driven story.

In the regression, firm-level financial ratios on the left-hand side include profitability risk as measured by $ProfitMargin$, operating risk as measured by operating cash flow divided by lagged total assets denoted as $OperatingCF$, long-term liquidity risk as measured by the solvency ratio denoted as $Solvency$, the innovation intensity as measured by the R&D expenses divided by lagged assets denoted as $R\&D$, the valuation risk as measured by Tobin's Q denoted as $Tobin'sQ$ and the dividend payout risk as measured by the payout ratios denoted as $Dividpayout$. The regression equation is as follows,

$$RISK_{i,t} = \alpha + GHG_{i,t} + After_t + GHG_{i,t} \times After_t + \gamma' X_{i,t-1} + \varepsilon_{i,t}, \tag{8}$$

where we include other fundamentals as control variables the same as in 4, and include industry fixed effects in the regressions. The sample period is from 2002 to 2021.

[Insert Table B4 near here]

Regression results are displayed in table B4. In columns 1 and 2 where the dependent variable is profit margin, the interaction term is significantly positive, with regression coefficients of 0.0069 and 0.0328, and t-statistics of 2.51 and 4.37, respectively. The interaction terms suggest that after the Paris Agreement, high-emission firms tend to have higher profitability ratios. Similarly, the high-emission firms tend to have higher operating cashflows indicating better operation ability, higher solvency rates, and higher dividend payout ratios. On the contrary, their R&D expenses are lower than low-emission firms. These contrasting results give rise to an alternate explanation: firms that spend more R&D expenses on clean technologies after the Paris Agreement face a new trade-off between operating efficiency, low-cost, or high business performance, and ESG performance. Low-carbon firms reduce their greenhouse gas emissions by adopting expensive low-carbon manufacturing technologies, thus earning lower returns for their shareholders.

However, this alternative hypothesis does not hamper the main message we would like to convey in the main empirical analysis, where investors' awareness is gradually

increased in the past few years, and ESG fund managers exploit this shift in preference by diverting their investments from the market portfolio towards the ESG portfolio, as documented in Pástor et al. (2021, 2022); van der Beck (2021).

This paper does not claim the shift in ESG preference is the only force that drives the high realized returns of low-carbon stocks over the past few years. Yet, we are mostly focusing on the empirical fact that stocks of firms with low carbon emissions tend to have higher realized returns after the Paris Agreement.

B.4. Carbon emission predicted with linear models

One potential concern with our machine-learning-based approach may be either a lack of transparency or overfitting. To overcome these obstacles, we use 5-fold cross-validation tests to determine the best iteration rounds in our main analysis. However, given our analysis is based on the premise that similar firms produce carbon emissions on a similar scale, we should expect this method to also apply to linear models.

As figure A5, we show that the machine learning model outperforms the linear model in terms of prediction accuracy, with an average out-of-sample R2 around 0.77 over 0.19. Preliminary results imply that carbon emission data predicted by XGBoost is more reliable.

To further our understanding of the basic premise and to examine the emission-return relationship in a more elucidatory fashion, we use carbon emission estimated with linear models as the dependent variable of interest, and examine whether and how carbon risk is priced in the cross-section of stock returns.

We follow equation 5 by running pooled OLS with three data samples. The first sample is from 2002 to 2016 (before the Paris Agreement), the second sample is from 2016 to 2021 (after the Paris Agreement), and the third sample is the whole sample period.

[Insert Table B5 near here]

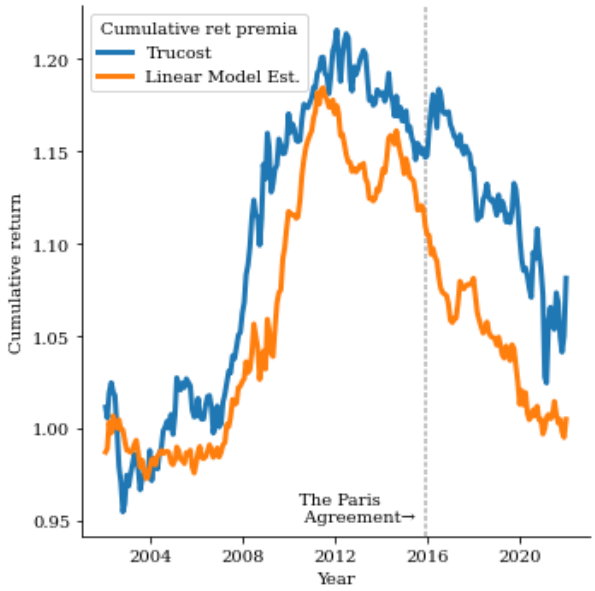
Regression results are displayed in table B5. The main difference between the data set estimated with XGBoost and the linear model lies within the sample period before the Paris Agreement. In panel A where we select a time period from 2002 to 2016, the regression coefficients in front of the logarithmic value of carbon emissions in columns 1 and 2 are 0.0139 and 0.0118 without and with industry fixed effects, and the t-statistics are 1.04 and 2.24, respectively. This contrasts with the significantly slightly negative emission-return relationship documented in table 11. Moreover, the economic magnitude is also much smaller, which suggests that the prediction by linear models just might be too noisy when there is a lot of missing data.

In panel B where we select the time period from 2016 to 2021, the emission-return relationship quickly reversed to become significantly negative, similar to the main analysis in table 5 and the supplementary evidence in table B2. The regression coefficients in columns 7 and 8 are -0.0591 and -0.0262, with t-statistics -2.60 and -0.78 respectively. We conjecture that there is not a huge difference between XGBoost's estimated results and the linear model's estimated result is that the Trucost database already provides ample estimation of carbon emissions (roughly over 3000 firms per year after 2016), and our estimation method only complements a rather smaller portion of the missing data (around 1000 firms). The main contribution of our estimation method lies before the Paris Agreement. The data set documents the dynamic shift in investors' ESG-related preferences.

We also report regression results with different emission variables including emission growth rate $GHGGR$ and emission intensity $GHGINTEN$. Similarly, carbon premia estimated with emission growth rate is volatile as we are estimating carbon emissions on a cross-sectional basis, and the premia estimated with emission intensity is consistently negative.

We also plot the cumulative emission premia from 2002 to 2021 following figure 5. We estimate the cumulative monthly return premia with the data set provided by Trucost and estimated by linear models. As shown in figure B1 Panel a, the cumulative premia decreased prior to 2008 and suddenly increased around 2009, then remained stable for around 6 years. On the contrary, the post Paris Agreement return premia in panel B remains positive after 2016. Therefore, this figure suggests that the dataset estimated with linear models tends to exacerbate the noise within the original patterns, thus making the predictions pretty unreliable. As a result, we argue that using the data set predicted by XGBoost might just as well be a better alternative.

[Insert Figure B1 near here]



Subfigure A: Cumulative premia

Subfigure B: Cumulative premia with industry FE

Fig. B1. Carbon cumulative return premia with the linear model approach. This figure plots the cumulative return premia estimated from the cross-sectional regressions of monthly returns from 5. The independent variable of interest is the natural logarithmic value of carbon emission, and the dependent variable is monthly stock returns. We adjust the magnitudes in terms of the unit standard deviation of the logarithmic emission at each cross-section following Bolton and Kacperczyk (2021a). We use either the Trucost sample or the one estimated by the linear model for estimation, and the sample period is from 2002 to 2021. In subfigure A, we do not include industry-fixed effects, whereas, in Panel B, we include industry-fixed effects. The first vertical dashed line denotes the ratification of the Paris Agreement, and the second dashed line denotes the start of the year 2021. For the Trucost sample, there are 374 distinct firms in 2021, as a result, we do not report cumulative return premia after 2021 for the blue line.

Table B1: Carbon emission and realized stock returns

Sample period	Panel A: 2002-2021 (Trucost sample)						Panel B: 2002-2021(XGBoost Sample)					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
LOGGHG	0.0098	0.0536**					-0.0239	-0.0226				
	(0.48)	(2.34)					(-1.56)	(-1.41)				
GHGGR			0.7061***	0.7412***					0.0262	0.0334*		
			(3.72)	(3.81)					(1.51)	(1.83)		
GHGINTEN					-0.0081	-0.0127					-0.0035***	-0.0035***
					(-0.76)	(-1.00)					(-5.33)	(-5.90)
LOGSIZE	0.2110	0.1799	0.3317**	0.3280**	0.2034	0.1959	0.1675*	0.1551*	0.2269**	0.2178**	0.1199	0.0957
	(1.31)	(1.15)	(2.06)	(2.03)	(1.30)	(1.30)	(1.73)	(1.68)	(2.38)	(2.35)	(1.28)	(1.10)
B2M	0.0051	-0.0071	0.0356	0.0401	0.0026	0.0004	0.0024	0.0040	0.0067	0.0064	-0.0066	-0.0117
	(0.15)	(-0.20)	(1.13)	(1.29)	(0.08)	(0.01)	(0.06)	(0.10)	(0.17)	(0.15)	(-0.17)	(-0.29)
LEVERAGE	0.0248	-0.3707	-0.0686	-0.2306	0.0132	-0.3141	0.2579	-0.0070	0.2717	-0.0531	0.1654	-0.1983
	(0.11)	(-1.40)	(-0.30)	(-0.86)	(0.06)	(-1.20)	(0.77)	(-0.02)	(0.83)	(-0.18)	(0.49)	(-0.67)
MOM	-0.1588	-0.1777	-0.1819	-0.2036	-0.1584	-0.1775	-0.3202***	-0.3291***	-0.3268***	-0.3364***	-0.3211***	-0.3291***
	(-1.47)	(-1.56)	(-1.56)	(-1.65)	(-1.47)	(-1.56)	(-3.21)	(-3.27)	(-3.01)	(-3.08)	(-3.17)	(-3.23)
INVEST2A	-4.1423***	-1.1284	-4.2961**	-1.3298	-4.0824***	-1.4080	-5.4358***	-4.0684***	-5.1627***	-3.5388***	-5.6029***	-4.2447***
	(-2.66)	(-0.94)	(-2.47)	(-1.13)	(-2.65)	(-1.17)	(-3.24)	(-3.27)	(-3.05)	(-2.81)	(-3.34)	(-3.33)
ROE	1.3541***	1.2599***	1.3136***	1.2813***	1.3487***	1.2730***	2.1391***	1.9564***	2.0618***	1.8760***	1.8997***	1.7943***
	(4.88)	(5.24)	(4.58)	(4.81)	(4.95)	(5.30)	(9.02)	(10.94)	(8.09)	(9.88)	(8.31)	(10.23)
HHI	0.0808	-1.3239	0.5122	-1.4447	0.0584	-1.2548	-1.0566	-2.5297	-1.1828	-3.2798	-0.9635	-2.4489
	(0.07)	(-0.77)	(0.54)	(-0.85)	(0.05)	(-0.73)	(-1.23)	(-1.22)	(-1.36)	(-1.48)	(-1.14)	(-1.20)
LOGPPE	-0.0698	-0.0277	-0.0822	-0.0303	-0.0500	0.0026	0.0974*	0.1349***	0.0576	0.1002*	0.0615	0.1148**
	(-1.30)	(-0.60)	(-1.41)	(-0.62)	(-0.88)	(0.06)	(1.69)	(2.72)	(0.98)	(1.84)	(1.08)	(2.32)
BETA	-0.5782***	-0.5715***	-0.5272**	-0.5147**	-0.5815***	-0.5690***	-0.5653***	-0.5511***	-0.5120**	-0.5021**	-0.5774***	-0.5538***
	(-2.66)	(-2.73)	(-2.30)	(-2.35)	(-2.65)	(-2.74)	(-2.80)	(-2.82)	(-2.55)	(-2.57)	(-2.88)	(-2.85)
VOLAT	0.3311***	0.3514***	0.3333***	0.3532***	0.3310***	0.3510***	0.3350***	0.3437***	0.3356***	0.3456***	0.3357***	0.3443***
	(3.49)	(3.64)	(3.37)	(3.51)	(3.49)	(3.64)	(6.70)	(6.96)	(6.28)	(6.53)	(6.66)	(6.93)
SALESGR	-0.5847**	-0.4992*	-0.6767**	-0.6188**	-0.5850**	-0.4968*	-0.5827***	-0.4879***	-0.6282***	-0.5318***	-0.6447***	-0.5734***
	(-2.04)	(-1.82)	(-2.23)	(-2.14)	(-2.05)	(-1.81)	(-3.68)	(-3.06)	(-3.60)	(-3.03)	(-4.22)	(-3.74)
EPSGR	0.0510*	0.0525*	0.0375	0.0379	0.0513*	0.0523*	0.0779***	0.0844***	0.0630***	0.0685***	0.0891***	0.0922***
	(1.75)	(1.71)	(1.33)	(1.30)	(1.75)	(1.72)	(3.48)	(3.78)	(2.78)	(3.08)	(3.81)	(3.98)
Const	T	T	T	T	T	T	T	T	T	T	T	T
Year-Mon FE	T	T	T	T	T	T	T	T	T	T	T	T
Ind FE		T		T		T		T		T		T
R2	0.22	0.22	0.23	0.23	0.22	0.22	0.16	0.16	0.16	0.16	0.16	0.16
N	332410	332410	299871	299871	332338	332338	764150	764150	680729	680729	760913	760913

This table examines the relation between carbon emission and stock realized returns with full sample data provided by Trucost and the one estimated by XGBoost. The dependent variable is the stock return of firm i in year t. The variables of interest are three measures of firms' carbon emissions, including the logarithmic value of carbon emission, emission growth rate, and emission intensity defined as carbon emission over firm sales. Other firm fundamentals include firm size, book-to-market ratio, leverage ratio, momentum, investment ratio, ROE, HHI index, Plant, property & equipment, Beta, return volatility, sales growth rate, and EPS growth rate. In panel A, we use the full sample from the Trucost data source. In panel B, we use the XGBoost estimated data panel. We control for year-month fixed effects and include industry-fixed effects separately in the regressions, and we cluster standard errors at the 2-digit GIC industry and year levels.

Table B2: Carbon emission and realized stock returns

Sample period	2016-2021 (After the Paris Agreement with Trucost sample)					
	(7)	(8)	(9)	(10)	(11)	(12)
LOGGHG	-0.0611**	0.0184				
	(-2.18)	(0.42)				
GHGGR			0.7999**	0.9668**		
			(2.03)	(2.17)		
GHGINTEN					-0.0330	-0.0299
					(-0.92)	(-1.02)
LOGSIZE	0.5433**	0.4548*	0.7225***	0.6344***	0.5171**	0.4543**
	(2.01)	(1.94)	(3.37)	(2.89)	(2.03)	(2.12)
B2M	0.0262	-0.0153	0.0744**	0.0450*	0.0216	-0.0132
	(0.49)	(-0.32)	(2.44)	(1.70)	(0.42)	(-0.30)
LEVERAGE	-0.0762	-1.1348***	-0.1272	-1.0350***	0.0163	-1.1131***
	(-0.21)	(-3.84)	(-0.43)	(-3.80)	(0.05)	(-3.79)
MOM	-0.2957*	-0.3599*	-0.3184*	-0.3826*	-0.2953*	-0.3598*
	(-1.75)	(-1.92)	(-1.69)	(-1.83)	(-1.73)	(-1.92)
INVEST2A	-3.6809*	-1.1691	-3.9566	-1.5685	-3.7556*	-1.2493
	(-1.71)	(-0.63)	(-1.54)	(-0.78)	(-1.85)	(-0.70)
ROE	1.9539***	1.6118***	1.8952***	1.6794***	1.9062***	1.6184***
	(5.79)	(5.68)	(5.30)	(5.68)	(5.88)	(5.84)
HHI	-0.8703	-22.7481	-0.2243	-12.3507	-1.0412	-22.6883
	(-0.40)	(-1.60)	(-0.13)	(-1.51)	(-0.51)	(-1.60)
LOGPPE	-0.1405	-0.0087	-0.2496***	-0.0528	-0.1695*	0.0081
	(-1.51)	(-0.16)	(-3.19)	(-0.66)	(-1.78)	(0.13)
BETA	-0.5702**	-0.4526*	-0.5158*	-0.4118	-0.5766**	-0.4491*
	(-2.00)	(-1.67)	(-1.68)	(-1.41)	(-1.98)	(-1.67)
VOLAT	0.4136***	0.4585***	0.4188***	0.4613***	0.4125***	0.4582***
	(3.21)	(3.54)	(3.10)	(3.34)	(3.19)	(3.53)
SALESGR	-0.8674*	-0.7174*	-0.8858*	-0.8201*	-0.8538*	-0.7179*
	(-1.78)	(-1.74)	(-1.66)	(-1.90)	(-1.75)	(-1.74)
EPSGR	0.0737*	0.0792*	0.0655	0.0679	0.0752*	0.0792*
	(1.83)	(1.82)	(1.52)	(1.44)	(1.78)	(1.83)
Const	T	T	T	T	T	T
Year-Mon FE	T	T	T	T	T	T
Ind FE		T		T		T
R2	0.22	0.23	0.23	0.24	0.23	0.23
N	160351	160351	142178	142178	160327	160327

This table examines the relation between carbon emission and stock realized returns with the Trucost data sample post the Paris Agreement. The dependent variable is the stock return of firm i in year t. The variables of interest are three measures of firms' carbon emissions, including the logarithmic value of carbon emission, emission growth rate, and emission intensity defined as carbon emission over firm sales. Other firm fundamentals include firm size, book-to-market ratio, leverage ratio, momentum, investment ratio, ROE, HHI index, Plant, property & equipment, Beta, return volatility, sales growth rate, and EPS growth rate. We control for year-month fixed effects and include industry-fixed effects separately in the regressions, and we cluster standard errors at the 2-digit GIC industry and year levels. The sample period is from 2016 to 2021.

Table B3: Returns and firm characteristics of different quintile portfolios

Panel A: Sorting results based on XGBoost estimated data

Portfolio	VW return	EW return	LOGGHG	GHGGR	GHGINTEN	SALE	AT	NCT	LOGSIZE	LEVERAGE	B2M	INVEST2A	ROE	HHI	LOGPPE	SALESGR	EPSGR
Lo	0.7909**	1.0858***	5.73	-0.07	1.00	2115.81	19052.24	1262.04	13.35	0.57	1.07	0.03	0.02	0.09	4.14	0.10	-0.01
	(2.28)	(2.90)															
2	0.7948**	1.1366**	10.58	0.18	16.34	4343.98	28683.47	2633.28	13.66	0.53	1.08	0.04	0.04	0.10	4.93	0.10	0.00
	(2.43)	(2.93)															
3	0.8365***	1.1146***	12.07	0.22	52.62	4001.82	17838.91	3213.46	13.29	0.51	0.91	0.04	-0.01	0.10	4.51	0.10	-0.06
	(2.79)	(2.75)															
4	0.5827**	0.8504**	13.04	0.39	81.02	3491.13	6621.45	3907.59	12.91	0.58	0.97	0.03	-0.01	0.09	4.17	0.09	-0.10
	(2.00)	(2.28)															
Hi	0.5660**	0.8970**	14.47	0.50	56.60	13561.63	21733.61	15808.27	14.22	0.63	1.43	0.05	0.06	0.10	6.29	0.09	0.00
	(2.14)	(2.50)															
Hi-Lo	-0.2248	-0.1889															
	(-1.30)	(-1.63)															

Panel B: Sorting results based on Trucost data

Portfolio	VW return	EW return	LOGGHG	GHGGR	GHGINTEN	SALE	AT	NCT	LOGSIZE	LEVERAGE	B2M	INVEST2A	ROE	HHI	LOGPPE	SALESGR	EPSGR
Lo	0.8431**	1.1558***	6.65	0.06	0.06	1998.78	15823.26	1537.70	13.96	0.62	1.02	0.02	0.01	0.09	4.18	0.12	0.04
	(2.45)	(3.04)															
2	0.9130***	1.2473***	9.02	0.09	0.14	5105.90	44352.70	2892.21	14.48	0.56	1.30	0.03	0.08	0.10	5.37	0.11	0.08
	(2.65)	(3.42)															
3	0.7336**	1.1071***	10.48	0.08	0.28	7484.42	55361.99	4752.33	14.79	0.56	1.25	0.04	0.10	0.10	6.20	0.08	0.03
	(2.28)	(2.91)															
4	0.7531***	0.9853***	11.97	0.08	0.90	13113.74	31351.79	11004.93	15.14	0.59	1.46	0.05	0.12	0.10	7.24	0.07	0.05
	(2.65)	(2.73)															
Hi	0.5964**	0.8254**	14.94	0.10	8.64	26088.84	37939.85	26966.58	15.41	0.64	2.23	0.07	0.10	0.09	8.54	0.06	0.04
	(2.24)	(2.29)															
Hi-Lo	-0.2467	-0.3304															
	(-1.20)	(-1.71)															

This table summarizes mean values of firm fundamentals of different portfolios sorted by three different measures of carbon emission, including the logarithmic value of carbon emission, emission growth rate, and emission intensity defined as carbon emission over firm sales. For each of the five portfolios sorted by different emission intensity variables, we report both equal-weighted and value-weighted portfolio returns. We also report other firm fundamentals, including sales, total assets, non-current assets, firm size, leverage ratio, book-to-market ratio, investment ratio, ROE, HHI index, Plant, property & equipment, sales growth rate, and EPS growth rate. In panel A, we report sorting results based on carbon emission data estimated by XGBoost, and in panel B, we report results estimated with Trucost data. The observation period is from 2002 to 2021, and we exclude financial firms with a 2-digit GIC industry classification of start with number 40.

Table B4: Carbon emission and other risk

	XGBoost sample											
Dep.Var	Profit margin		Operating CF		R&D		Solvency		Tobin's Q		Divid payout	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
LOGGHG	-0.0013	-0.0012	0.0001	-0.0004	0.0001	0.0003**	0.0382	-1.0861***	0.0023	-0.0120***	-0.0012	-0.0025*
	(-0.88)	(-0.86)	(0.58)	(-1.57)	(0.67)	(2.37)	(0.17)	(-3.68)	(1.07)	(-4.57)	(-0.61)	(-1.77)
LOGGHG*POST	0.0069**	0.0328***	0.0003	0.0023***	-0.0002	-0.0013***	0.3376	0.8784*	-0.0010	-0.0156**	0.0055**	0.0065***
	(2.51)	(4.37)	(0.56)	(3.20)	(-1.25)	(-4.82)	(0.82)	(1.84)	(-0.22)	(-2.57)	(2.06)	(2.79)
POST	-0.0676**	-0.4273***	-0.0205***	-0.0522***	0.0022	0.0151***	-10.2744*	-17.7466***	-0.0257	0.2418**	-0.0758**	-0.1109***
	(-2.21)	(-5.93)	(-2.84)	(-5.84)	(0.99)	(4.56)	(-1.80)	(-2.96)	(-0.34)	(2.59)	(-2.26)	(-3.87)
LOGSIZE	-0.0256**	-0.0259***	0.0107***	0.0037***	0.0008	0.0075***	7.2689***	4.4974***	0.7102***	0.4636***	0.0592***	0.0798***
	(-2.50)	(-3.35)	(5.00)	(3.47)	(1.16)	(13.27)	(4.78)	(5.10)	(26.94)	(21.52)	(4.94)	(12.71)
B2M	-0.0015	-0.0069**	-0.0010	-0.0011**	-0.0007**	0.0015***	-0.2542	0.8241	0.0773***	0.0660***	-0.0033	0.0065*
	(-0.27)	(-2.19)	(-0.84)	(-2.06)	(-2.42)	(5.14)	(-0.39)	(1.60)	(3.71)	(8.21)	(-0.52)	(1.73)
ROE	0.7208***	1.2703***	0.1366***	0.2586***	-0.0105***	-0.0487***	11.6931***	45.0067***	-0.1209***	-0.0241	0.2418***	0.3837***
	(24.32)	(25.75)	(23.98)	(31.87)	(-6.40)	(-14.82)	(4.98)	(15.81)	(-3.81)	(-0.54)	(10.35)	(19.95)
LEVERAGE	0.0984	0.3036***	-0.0147*	-0.0147*	0.0099***	0.0001	-92.3579***	-111.3659***	0.7219***	0.0209	0.1975***	0.0909***
	(1.55)	(4.84)	(-1.88)	(-1.88)	(3.81)	(0.02)	(-10.16)	(-13.39)	(6.80)	(0.27)	(2.96)	(3.38)
INVEST2A	-0.0954	-0.0657	0.1883***	0.3349***	0.0436***	0.0569***	54.8388***	97.2384***	3.1299***	5.4596***	0.2873	-0.5441***
	(-0.68)	(-0.44)	(6.57)	(10.02)	(5.19)	(4.83)	(3.50)	(5.49)	(10.90)	(17.34)	(1.35)	(-3.35)
HHI	-0.0080	-0.0517	-0.0686**	-0.0348	-0.0045	-0.0061	-20.4462	-14.0915	-0.0689	0.8558***	0.1076	-0.0102
	(-0.05)	(-0.37)	(-2.46)	(-1.56)	(-0.56)	(-0.61)	(-0.91)	(-0.69)	(-0.19)	(2.73)	(0.47)	(-0.05)
LOGPPE	0.0762***	0.0571***	0.0033*	0.0098***	-0.0060***	-0.0096***	-3.7427***	-0.7655	-0.5251***	-0.3681***	0.0120	-0.0080
	(5.34)	(8.17)	(1.71)	(8.75)	(-5.98)	(-17.16)	(-2.83)	(-0.88)	(-15.92)	(-21.38)	(1.05)	(-1.34)
SALESGR	0.3490***	0.1663***	0.0640***	0.0394***	0.0096***	0.0143***	12.8731***	8.3527***	0.1020***	0.3035***	-0.1236***	-0.1941***
	(9.86)	(5.42)	(19.13)	(7.98)	(9.40)	(10.36)	(7.04)	(5.06)	(3.55)	(11.48)	(-5.53)	(-8.34)
EPSGR	0.0078***	-0.0131***	0.0013***	-0.0032***	0.0002**	0.0017***	1.5774***	0.6618***	0.0055**	0.0063	0.0251***	0.0234***
	(3.45)	(-4.51)	(3.87)	(-5.03)	(2.06)	(8.85)	(5.88)	(3.07)	(2.23)	(1.39)	(6.81)	(7.69)
Const	T	T	T	T	T	T	T	T	T	T	T	T
Ind FE		T		T		T		T		T		T
R2	0.22	0.35	0.39	0.60	0.10	0.24	0.08	0.16	0.08	0.28	0.17	0.19
N	67720	67720	67356	67356	67911	67911	51379	51379	67873	67873	63766	63766

This table examines the relationship between carbon emission and other firm-level risks on the firm-year level. The dependent variables are a host of firm-level risks including profitability risk as measured by profit margin, operating risk as measured by operating cash flow divided by lagged total assets, long-term liquidity risk as measured by the solvency ratio, the innovation intensity as measured by the R&D expenses divided by lagged assets, the valuation risk as measured by the Tobin's Q, and the dividend payout risk as measured by the payout ratios. The independent variables include the logarithmic value of firms' carbon emissions, a dummy that indicates after the Paris Agreement, and their interaction term. We also include other fundamentals as control variables the same as in 4, and control for industry fixed effects in the regressions. The sample period is from 2002 to 2021.

Table B5: Carbon emission and realized stock returns: emission estimated with linear models

Carbon emissions estimated by linear models												
Sample period	Panel A: 2002-2016 (before the Paris agreement)						Panel B: 2016-2021 (after the Paris agreement)					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
LOGGHG	0.0139	0.0118**					-0.0591**	-0.0262				
	(1.04)	(2.24)					(-2.60)	(-0.78)				
GHGGR			0.0305	0.0306					-0.0170	0.0167		
			(1.33)	(1.36)					(-0.58)	(0.55)		
GHGINTEN					-0.0015***	-0.0016***					-0.0041***	-0.0035***
					(-5.36)	(-5.71)					(-4.52)	(-4.03)
LOGSIZE	0.0891	0.0685	0.1262	0.1505	0.0456	0.0382	0.3856*	0.3621**	0.4863**	0.4062**	0.3114	0.2859*
	(0.92)	(0.75)	(1.14)	(1.27)	(0.47)	(0.42)	(1.94)	(2.06)	(2.25)	(2.13)	(1.60)	(1.74)
B2M	0.0301	0.0358	0.0121	0.0341	0.0145	0.0275	-0.0551	-0.0837	-0.0367	-0.0710	-0.0752	-0.1060
	(0.74)	(0.82)	(0.30)	(0.74)	(0.39)	(0.66)	(-0.68)	(-1.08)	(-0.40)	(-0.87)	(-0.89)	(-1.31)
LEVERAGE	0.2170	0.1938	0.0838	-0.0087	-0.0533	0.0356	0.2148	-0.9942**	0.0305	-0.8305*	0.1230	-1.1062***
	(0.62)	(0.52)	(0.23)	(-0.03)	(-0.13)	(0.10)	(0.40)	(-2.31)	(0.05)	(-1.80)	(0.24)	(-2.71)
MOM	-0.2312**	-0.2378**	-0.2118	-0.2211*	-0.2321**	-0.2396**	-0.5255***	-0.5736***	-0.5399***	-0.5891***	-0.5200***	-0.5648***
	(-2.15)	(-2.20)	(-1.65)	(-1.72)	(-2.16)	(-2.21)	(-2.89)	(-3.09)	(-2.90)	(-3.08)	(-2.73)	(-2.93)
INVEST2A	-5.1072***	-4.6327***	-5.4753***	-4.6627***	-4.5399**	-4.5971***	-7.2148***	-5.2973*	-6.2520**	-5.0361	-7.6256***	-5.5896*
	(-2.81)	(-4.33)	(-2.96)	(-4.20)	(-2.45)	(-4.16)	(-3.10)	(-1.73)	(-2.36)	(-1.58)	(-3.38)	(-1.93)
ROE	1.7748***	1.7032***	1.5637***	1.4270***	1.6663***	1.6431***	3.0178***	2.4532***	2.5609***	2.0571***	2.7615***	2.3126***
	(7.26)	(7.61)	(8.03)	(8.43)	(6.67)	(7.12)	(7.73)	(13.77)	(5.75)	(9.73)	(7.59)	(14.30)
HHI	-0.6750	-2.4455	-0.0517	3.9985***	-0.7187	-2.3683	-1.8100	-8.4692	-0.8461	-6.2778	-2.0304	-7.8137
	(-1.33)	(-1.11)	(-0.08)	(2.71)	(-1.40)	(-1.07)	(-0.73)	(-0.50)	(-0.34)	(-0.36)	(-0.83)	(-0.46)
LOGPPE	0.1115**	0.1378***	0.1206	0.1214*	0.1384**	0.1325**	0.0458	0.1470	-0.0521	0.1128	-0.0192	0.1220
	(2.07)	(2.63)	(1.59)	(1.86)	(2.16)	(2.59)	(0.35)	(1.34)	(-0.33)	(0.91)	(-0.16)	(1.19)
BETA	-0.6260***	-0.6053***	-0.6105***	-0.5910***	-0.6305***	-0.6033***	-0.3388	-0.2110	-0.2773	-0.1426	-0.3467	-0.2197
	(-2.72)	(-2.71)	(-2.79)	(-2.74)	(-2.74)	(-2.69)	(-1.20)	(-0.70)	(-1.01)	(-0.49)	(-1.23)	(-0.73)
VOLAT	0.3134***	0.3189***	0.3035***	0.3141***	0.3165***	0.3215***	0.3975***	0.4242***	0.4004***	0.4260***	0.3962***	0.4213***
	(4.62)	(4.80)	(3.79)	(3.94)	(4.63)	(4.80)	(5.22)	(5.52)	(5.45)	(5.68)	(5.05)	(5.34)
SALESGR	-0.5262***	-0.4875***	-0.5895***	-0.5224***	-0.5481***	-0.5416***	-0.7409*	-0.6313*	-0.8416*	-0.7201*	-0.7828**	-0.6911**
	(-3.54)	(-3.13)	(-3.53)	(-3.10)	(-3.78)	(-3.65)	(-1.95)	(-1.91)	(-1.82)	(-1.83)	(-2.03)	(-2.03)
EPSGR	0.0965***	0.0998***	0.0146	0.0202	0.1026***	0.1037***	0.0307	0.0352	0.0238	0.0290	0.0410	0.0396
	(3.70)	(3.76)	(0.57)	(0.86)	(3.81)	(3.80)	(0.72)	(0.89)	(0.66)	(0.89)	(1.02)	(1.07)
Const	T	T	T	T	T	T	T	T	T	T	T	T
Year-Mon FE	T	T	T	T	T	T	T	T	T	T	T	T
Ind FE		T		T		T		T		T		T
R2	0.16	0.16	0.18	0.18	0.16	0.16	0.16	0.17	0.17	0.17	0.17	0.17
N	541623	541623	342682	342682	539640	539640	231285	231285	199564	199564	229913	229913

This table examines the relation between carbon emission and stock realized returns with data samples estimated with linear models. The dependent variable is the stock return of firm i in year t. The variables of interest are three measures of firms' carbon emissions, including the logarithmic value of carbon emission, emission growth rate, and emission intensity defined as carbon emission over firm sales. Other firm fundamentals include firm size, book-to-market ratio, leverage ratio, momentum, investment ratio, ROE, HHI index, Plant, property & equipment, Beta, return volatility, sales growth rate, and EPS growth rate. We predict carbon emission with linear models using the same covariates as in the XGBoost model. We use three different sample periods for empirical estimation. The first sample period is from 2002 to 2016 (before the Paris Agreement), whereas the second sample is after the Paris Agreement. The third sample uses the whole period. We add the same control variables as well as fixed effects following regression 5 in the main analyses, and we cluster standard errors at the 2-digit GIC industry and year levels.

Appendix C. An overview of the states' decarbonization targets

States in the US have been actively engaged in promoting low-carbon styles. A wide range of policies has been adopted at the state and regional levels to reduce greenhouse gas emissions, develop clean energy resources, promote alternative fuel vehicles, and promote more energy-efficient buildings and appliances, among other things. According to the C2ES database [3], twenty-four states plus the District of Columbia have adopted specific greenhouse gas emissions targets by the end of 2022. While each state has adopted a target and baseline year that suits its circumstances, the prevalence of these targets shows the widespread support for climate action. Figure C1 illustrates various types of carbon emission targets announced by different states.

[Insert Figure C1 near here]

We present regulation details by each state in table C1. For example, California set an executive target in 2018 to reach net-zero carbon dioxide emissions by 2045. The state previously set an executive target in 2005 to reduce GHG emissions to 80% below 1990 levels by 2050. The state enacted a statutory target in 2006 to reduce GHG emissions to 1990 levels by 2020; in 2016, it set a statutory target to reduce GHG emissions 40% below 1990 levels by 2030.

[Insert Table C1 near here]

[3] Detailed information can be accessed via https://www.c2es.org/content/state-climate-policy/, which describes greenhouse gas emission targets, state climate action plans, carbon pricing, and other related policies.

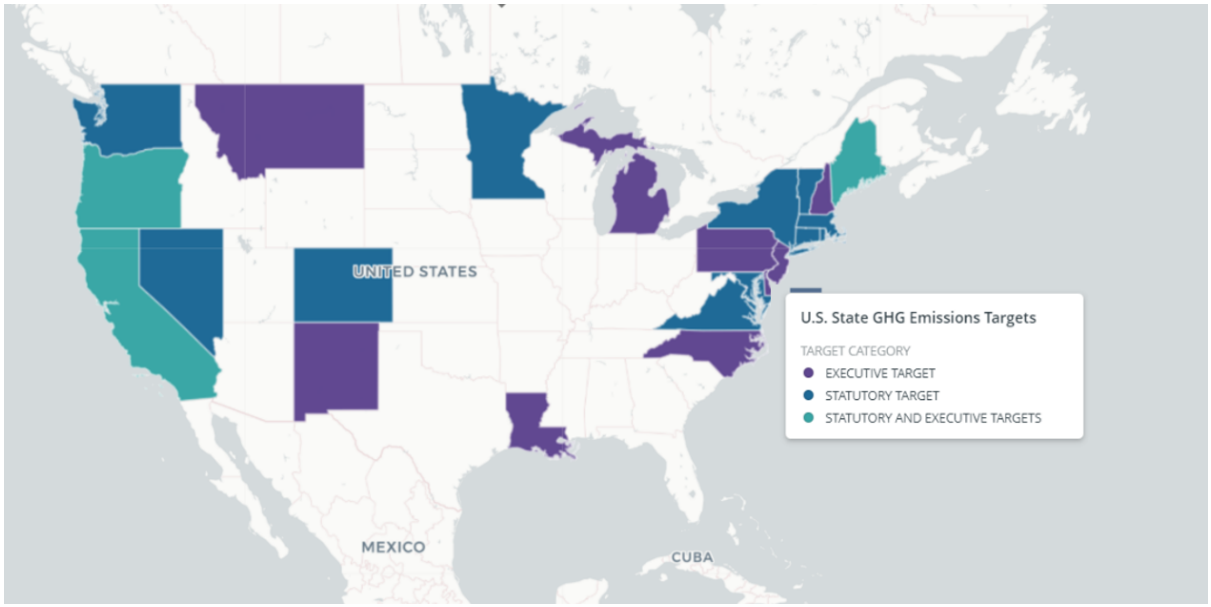


Fig. C1. States that announce emission reduction policies by the end of 2022. This figure is obtained from C2ES with minor revisions. The purple states are states that have announced executive targets for carbon emissions. The blue states are states that have set statutory targets for carbon emissions. The green states announce both types of targets.

Table C1: Emission target announcement details

State name	Earliest announcement year	Emission target details
CA California	2005	California set an executive target in 2018 to reach net-zero carbon dioxide emissions by 2045. The state previously set an executive target in 2005 to reduce GHG emissions 80% below 1990 levels by 2050. The state enacted a statutory target in 2006 to reduce GHG emissions to 1990 levels by 2020; in 2016, it set a statutory target to reduce GHG emissions 40% below 1990 levels by 2030.
CO Colorado	2019	Colorado enacted statutory targets in 2019 to reduce GHG emissions 26% by 2025, 50% by 2030, and 90% by 2050, all compared to 2005 levels.
CT Connecticut	2018	Connecticut enacted a statutory target in 2018 to reduce GHG emissions 45% below 2001 levels by 2030 and 80% by 2050. Previously, the state enacted statutory targets in 2008 to reduce GHG emissions at least 10% below 1990 levels by 2020 and 80% below 2001 levels by 2050.
DC District of Columbia	2017	The District of Columbia set executive targets in 2017 to reduce GHG emissions 50% below 2006 levels by 2032 and 80% below 2006 levels by 2050. In 2017, it also set a target to reach GHG emissions neutrality by 2050.
IL Illinois	2019	Illinois set an executive target in 2019 to reduce GHG emissions 26–28% below 2005 levels by 2025.
LA Louisiana	2020	Louisiana set executive targets in 2020 to reduce net GHG emissions 26–28% by 2025 and 40–50% by 2030, compared to 2005 levels. The targets also aim for net-zero GHG emissions by 2050.
MA Massachusetts	2021	Massachusetts enacted a statutory target in 2021 to reduce GHG emissions 85% below 1990 levels by 2050, with the goal of achieving net-zero emissions by 2050 and an interim target of reducing emissions at least 75% below 1990 levels by 2040. The state also set interim targets in 2022 to reduce GHG emissions 33% by 2025 and 50% by 2030 from 1990 levels.
MD Maryland	2016	Maryland enacted a statutory target in 2016 to reduce GHG emissions 40% below 2006 levels by 2030.
ME Maine	2019	In 2019, Maine set an executive target to achieve net-zero GHG emissions by 2050, and enacted statutory targets to reduce GHG emissions 45% below 1990 levels by 2030 and 80% below 1990 levels by 2050.
MN Minnesota	2007	Minnesota enacted statutory targets in 2007 to reduce GHG emissions 30% below 2005 levels by 2025 and 80% below 2005 levels by 2050.
MT Montana	2019	Montana set an executive target in 2019 to achieve economy-wide GHG neutrality with no set target year; in 2020, the state set the target year to reach economy-wide GHG neutrality between 2045–50.
NC North Carolina	2022	North Carolina set an executive target in 2022 to reduce GHG emissions 50% below 2005 levels by 2030, and to reach net-zero GHG emissions as soon as possible but no later than 2050.
NJ New Jersey	2007	New Jersey enacted statutory targets in 2007 to reduce GHG emissions to 1990 levels by 2020 and 80% below 2006 levels by 2050. The state also set an executive interim target in 2021 to reduce emissions 50% below 2006 levels by 2030.
NM New Mexico	2019	New Mexico set an executive target in 2019 to reduce GHG emissions 45% below 2005 levels by 2030.
NV Nevada	2019	Nevada enacted statutory targets in 2019 to reduce GHG emissions 28% by 2025 and 45% by 2030, compared to 2005 levels, and reach zero or near-zero emissions by 2050.
NY New York	2019	New York enacted statutory targets in 2019 to reduce GHG emissions 40% below 1990 levels by 2030 and at least 85% below 1990 levels by 2050. The targets also aim for net-zero GHG emissions by 2050.
OR Oregon	2007	Oregon set executive targets in 2020 to reduce GHG emissions 45% below 1990 levels by 2035 and 80% below 1990 levels by 2050. The state also enacted statutory targets in 2007 to reduce emissions 10% below 1990 levels by 2020 and 75% below 1990 levels by 2050.
PA Pennsylvania	2019	Pennsylvania set executive targets in 2019 to reduce GHG emissions 26% below 2005 levels by 2025, and 80% below 2005 levels by 2050.
RI Rhode Island	2021	Rhode Island enacted statutory targets in 2021 to reduce GHG emissions 10% by 2020, 45% by 2035, and 80% by 2040, all compared to 1990 levels. The targets also aim for net-zero GHG emissions by 2050.
VA Virginia	2020	Virginia enacted a statutory target in 2020 to achieve net-zero GHG emissions across all sectors by 2045.
VT Vermont	2020	Vermont enacted statutory targets in 2020 to reduce GHG emissions 26% below 2005 emissions by 2025, 40% below 1990 levels by 2030, and 80% below 1990 levels by 2050.
WA Washington	2020	Washington enacted statutory targets in 2020 to reduce GHG emissions 45% by 2030, 70% by 2040, and 95% by 2050, all compared to 1990 levels. The targets also aim for net-zero GHG emissions by 2050.
WI Wisconsin	2019	Wisconsin set an executive target in 2019 to reduce GHG levels by 26–28% below 2005 levels by 2025.