The low-carbon premium: a machine learning approach

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

March 30, 2023

Abstract

Firms that share similar business structures produce carbon emissions on a similar scale. In this paper, we estimate a large panel of carbon emission data by US firms with a novel approach. We use a measure called *Cosine Similarity* to proxy for firms' business similarity from 10-K filings. Then, we apply a machine learning algorithm known as *XGBoost* to estimate Scope 1 carbon emissions of listed firms from 2002 to 2021. Our estimated dataset has on average 4111 firms per year, which outnumbers all the other data sources and has high credibility. Based on this data set, we examine firms' carbon risk pricing in the US equity market. The result supports a low-carbon premium hypothesis, in which firms that emit less carbon dioxide significantly perform better, and this effect is more pronounced after the Paris Agreement at the end of 2015, implying investors' preference for carbon risk only became prominent in recent years. Overall, this paper reveals the possible sampling bias in previous research and provides researchers with a different approach to understanding climate finance.

JEL classification: G12, G23, G30.

Keywords: Carbon emission, Asset pricing, Cosine Similarity, XGBoost.

*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 Finvolution Group Ltd. for their invaluable technical support. We benefited from extensive discussions with Xiaomeng Lu and Shumiao Ouyang. All errors remain ours.

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, expand their coverage after the Paris agreement which was announced 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 which have disclosed carbon emissions based on the business similarity score and other firm fundamentals[1]. We make a prediction of a large data panel from 2002 to 2021 of US-listed stocks with this methodology and find convincing evidence supporting the existence of a low-carbon premium, and this low-carbon return premium is more pronounced after the Paris agreement.

Following Hoberg and Phillp's pioneering research on business similarity (Hoberg and Phillips, 2010, 2016) and Cohen et al. (2020), 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).

Next, we feed the cosine similarity scores, 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 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

[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 to predict carbon emissions in our case.

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 prior to 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 computed is robust, and we designed 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 33.83% of its carbon emission as compared to the control groups.

We also use a transition matrix to examine the persistence of carbon emissions. It is intuitive that 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 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 logarithm value of its carbon emissions, along with other firm fundamentals. Regression results suggest that the higher the carbon emission of a firm, the lower 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.

[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 emission in 2021.

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 prior to 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 which spans from 2002 to 2017, we find there exhibits a significantly positive carbon premium, and the effect is more pronounced once we control for industry fixed effect. However, when we use the data sample estimated by XGBoost and enlarge the database to include more firms and longer observation periods, the previously documented premium becomes insignificant. When we further narrow down the sample period from 2016 to 2021 after the Paris agreement, the premium significantly turns negative. The regression coefficient is -0.0888 and -0.0536 without and with the industry fixed effects, with t-statistics -3.32 and -2.12, respectively. The effect is more pronounced 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 are diverting their position towards less-carbon industries over the past few years.

We further test the time-varying carbon risk premium and the impact of the well-known Paris agreement with additional tests. Regression results 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 becomes most significant in 2020. Regressions with common risk factors show that the shift in preference is not driven by risk preferences like size, earnings, margin investments, or liquidity, instead, this change seems to be purely driven by ESG-related issues. This relationship between asset prices and carbon emission, if not causal, is negatively correlated at least. This shift is not likely to be driven by improvements in firm fundamentals but by investors' attention (Alekseev et al., 2022; Choi et al., 2020a,b). 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.

Admittedly, the relationship 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 returns. And yet, we argue that this paper at least provides researchers with a potential method to examine the pricing of carbon emissions prior to 2016, and also an interesting 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 into two main findings. First, when we include more listed firms in the sample, the positive relationship between carbon emissions and stock returns becomes insignificant from 2002 to 2021, revealing the self-selection bias in the original data set provided by Trucost. 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. Second, carbon risk only emerged after 2012 and becomes increasingly prominent after the Paris agreement, implying a positive shock for the ESG-related preference in the past decades. We expect to observe the low-carbon premium in the next few years, and the emission-return relationship will become positive again. It should be noted that we are not challenging the positive emission-return relationship documented in previous research. Their insightful analyses just may be limited to a lack of data and simply neglects the evolving interest in sustainable investment in recent years.

To the best of our knowledge, this is the first paper that tries to predict carbon emissions with machine learning algorithms. This novel approach proves to be reliable and efficient in predicting carbon emissions. Besides carbon emissions, researchers could also use this approach to predict ESG scores, patents, and other important variables. 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 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. This is also the first paper that examines the time-varying risk premium for carbon emissions as we have more emission data to explore the pricing of carbon emissions over time.

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 low-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. (2020) 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.

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). 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). 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). Overall, the literature gives inconclusive results and demands more empirical validation.

3. Greenhouse Gas data estimation

3.1. Data and variables

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 select 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 in order that we can filter noise, as many firm pairs have similarity scores close to zero. We present summary statistics of the similarity scores by year in table 1. We surely can expand the selection threshold from 20 to 30 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.

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

3.2. GHG prediction with XGBoost

We use a regression tree method, XGBoost, for predicting greenhouse gas emissions. 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.

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 non-disclosure firm and the disclosure 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 duplicate the observation for each similarity pair by switching the disclosure and non-disclosure firms. As a result, we have 229396 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 145576 observations in the training set and 83820 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 353070 observations, where we can use disclosed emission, firm similarity scores as well as other firm fundamentals to predict the carbon emission of the non-disclosure 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 1 near here]

We present an illustrative example of model training in figure 1, 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 2, 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 2, 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, 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 2 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 3. 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. Our method is able to 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 3. 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. 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. Then we apply the algorithm to the prediction set, where only one side of the similarity pair has disclosed carbon emis-

sions. 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 prediction, so as to enlarge and balance the dataset at full potential.

[Insert Figure 3 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 clearly seen that our data outnumber data provided by other vendors, especially prior to 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 9.49 and 9.36, median logarithmic emission of 9.44 and 9.76, respectively. The standard deviation of our dataset is slightly larger than the original data, as our database contains much more firms. Prior to 2016, the Trucost database only includes emissions of large firms or heavy pollution firms, which drastically increases 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 log emissions. The highest emission industry is utilities, followed by materials and energy. These three industries tend to emit massive greenhouse gas 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, while 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 purely 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 logarithm 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 distribution of the carbon emission generally is skewed to the right. In panel B, we report the correlation matrix between carbon emission between other firm fundamentals. 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 online appendices which is not reported here due to page limits.

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 Republic 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 depends on high-carbon petrol and chemical firms. Our regression formula is as follows.

$$LOGGHG_{i,t} = \text{ RegulatoryShock }_{i,t} + Controls_{i,t} + \mu_j + \lambda_s + \varepsilon_{i,t} \tag{2}$$

The dependent variable $LOGGHG$ is the logarithmic value of the firm's scope 1 greenhouse gas emission. The independent variable $Shock_{i,t}$ 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 logarithm 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 33.83% 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 measure 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 10.55% 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.

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 emission is 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 70.94%. For firms assigned to group 5 (the highest carbon emission group) at year 0, the probability that it remained in group 1 Is 83.34%. 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 would 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. 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. 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.

$$\text{Inclusion}_{\ i,t} = LOGGHG_{i,t} + Controls_{i,t} + \mu_j + \delta_t + \varepsilon_{i,t} \tag{3}$$

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 -0.1336 with a t-statistic of -3.34, 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 -0.0025 in the fifth column, with a t-statistic of -1.78. 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.

$$\text{GHG}_{\ i,t} = Controls_{i,t} + \mu_j + \delta_t + \varepsilon_{i,t} \tag{4}$$

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 $Control_{i,t}$ 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. 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 0.3414 and 0.3542 for both samples, with t-stats 14.06 and 11.23, 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 4 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.

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. r

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 premiums 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:

$$Ret_{i,t} = Const + GHG_{i,t} + \text{ Controls }_{i,t-1} + \delta_t + \mu_j + \varepsilon_{i,t}, \tag{5}$$

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 2016, and has 215808 observations. The second sample is emission data predicted from the XGBoost model with a full sample period from 2002 to 2021. It has 764150 observations. The third sample period narrows down the second sample from the start of 2016 to 2021. We choose 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 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 logarithm of carbon emissions, the coefficient is 0.0330 with t-statistics of 1.89, suggesting that higher emission firms earn higher stock returns. 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 significant, and emission growth rates are not. However, when we substitute the original data set from the one provided by Trucost as the one estimated with the XGBoost model, the baseline estimation yields different results, as the regression coefficient becomes insignificant in columns 7 and 8. Moreover, the emission-return relationship turns significantly negative for emission intensity in columns 11 and 12. In the last sample, where we narrow down the observation period into 2016 to 2021, the baseline estimation yields completely different results than the positive relationship in the first and second columns. The regression coefficients are -0.088 and -0.0536, with t-statistics of -3.32 and -2.12, 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 the last few 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 estimated with emission intensity which is normalized with different firm fundamentals, including total assets, non-current assets, Plant, property & equipment, and firm's market capitalizations, 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 and resulting in large deviations in the data.

[Insert Table 11 near here]

To examine the time-varying emission-return relationship, we perform regression in 6 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 are indefinite prior to 2015, and they turned consistently negative after 2015. 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 begin after 2012. 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]

We further explore the relationship between carbon emissions and stock returns before and after the Paris agreement. To do so, we first sort firms into five quintiles from low to high based on three measures of carbon emissions. Then, we keep the lowest quintile and the highest quintile firms and define a dummy variable $HIGHG$ to indicate whether the firm is a high-emission firm based on the emission group it belongs to. We interact the dummy variable with a time dummy that indicates after the Paris agreement. We estimate the following regression:

$$Ret_{i,t} = Const + HiEmi_{i,t} + After_t + HiEmi_{i,t} \times After_t + \text{ Controls }_{i,t-1} + FE + \varepsilon_{i,t}, \quad (6)$$

where the dependent variable is monthly stock returns, and the independent variable includes two dummy variables and their interactions, plus other firm financial variables. We control for year-month fixed effects and industries effects in different regressions as denoted by FE and cluster standard errors on firm and year levels. The sample we use is the data set estimated by the XGBoost algorithm from 2002 to 2021. To rule out potential concerns that the coarse industry classification methodology main results in our results, we exclude salient industries with 6-digit GIC code 101020 (Oil, Gas & Consumable Fuels), 551020 (Gas Utilities), and industries with 2-digit code 20 (Transportation). Regression results are displayed in table 13, where the coefficients in front of the interaction term in baseline regressions are all significantly negative, supporting our hypothesis that investors divert their portfolio toward less carbon-intensive firms after the Paris agreement. Moreover, this effect is slightly more substantial when we exclude salient industries. In columns 3 to 4 and columns 9 to 10, the interaction is not significant, whereas it is again negatively significant for the emission intensity variable.

Different results estimated with different measures suggest that following Aswani et al. (2022), it is delicate for academic researchers and industry practitioners to choose which emission measure to use. In figure 4, we present more illustrative evidence supporting the low-carbon premium hypothesis after the Paris agreement. 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 4, the portfolio returns appear to be positive prior to 2012 for value-weighted returns, and the relationship soon reversed afterward. This negative emission-return relationship becomes increasingly more significant after the end of 2015 and peaked at the end of 2020. We show that our portfolio sorting results also hold using data provided by Trucost in the appendices. Overall, reveals the sample bias that may exist in previous research. The carbon premium did exist prior to the Paris agreement, but investors' preferences significantly changed in the last decade.

[Insert Figure 4 near here]

4.2. The low-carbon premium and common risk factors

Finally, we examine the relationship between low-carbon premiums and risk factors. We estimate a time-series regression model using monthly premium, which is estimated from monthly cross-sectional regression in 6. We run the following regression:

$$RiskPrem_t = \alpha + Factor_{i,t} + \varepsilon_t \tag{7}$$

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 low-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 14 near here]

Regression results in columns 1 and 2 of table 14 are insignificant, 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 significant in columns 7 and 8 after we narrow down the sample period from 2016 to 2021. The estimated alphas in both columns are -0.0957 and -0.0796 with-and-without controlling for common risk factors, with t-statistics of -8.02 and -4.85, respectively. Interestingly, controlling for risk factors even makes the risk premium stronger and more significant. Besides, both the economic and statistical significance are stronger when we control for risk factors. In the next set of columns 9 and 10, we report risk premiums 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.

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 prior to 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 and conduct asset pricing tests, and find that, on average, stocks of high-emission firms consistently underperform stocks of low-emission firms. This emission-return relationship only emerged after 2012 and was ambivalent before 2012, and it becomes increasingly prominent after 2016. Common risk factors cannot explain the low-carbon risk premium, 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 expect this phenomenon could be persistent in the next few years, and finally, in a new equilibrium, the emission-return relationship will reverse.

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 social-economic impact on its surroundings beyond finance. Our prediction method based on business similarity networks can also be adopted with

other firm network data, so as to estimate other types of datasets 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

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.

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

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.

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., Malloy, C., Nguyen, Q., 2020. 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.

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 .

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

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.

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

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

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.

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.

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.

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

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.

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) .

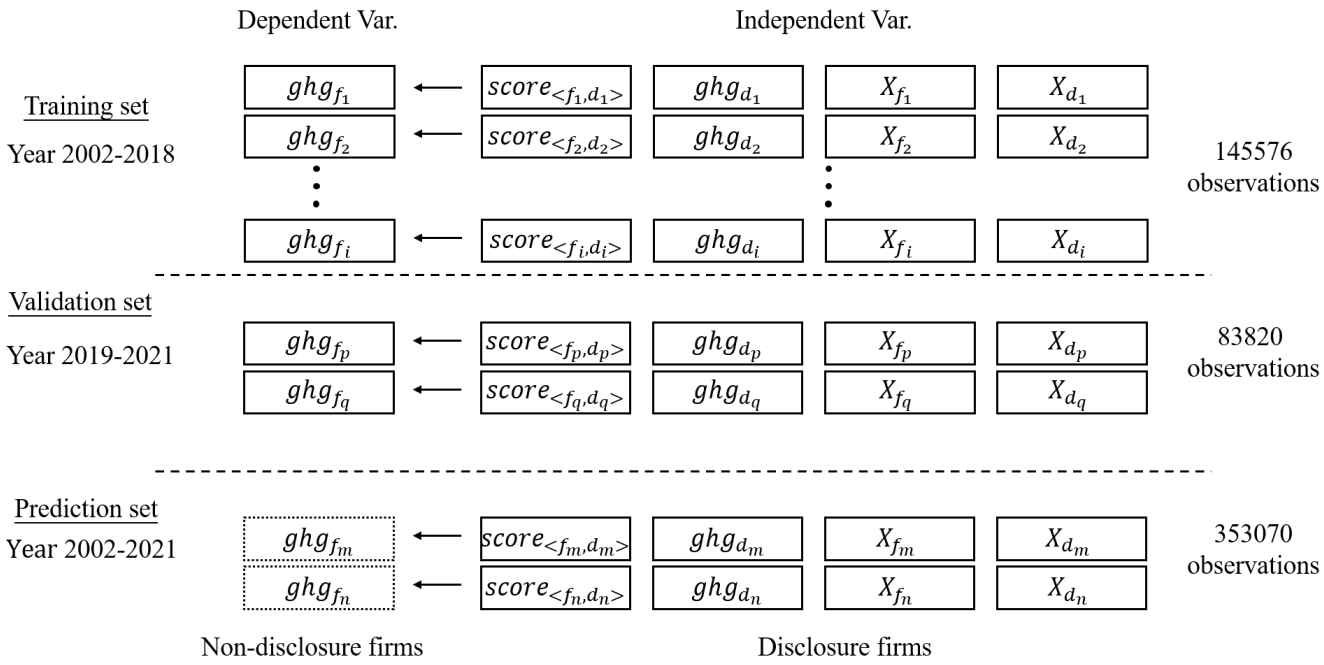
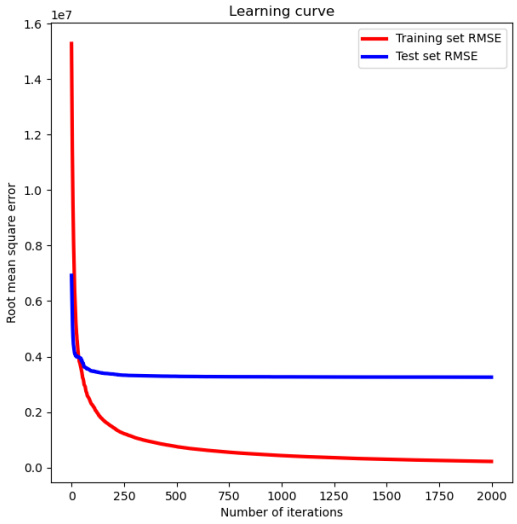
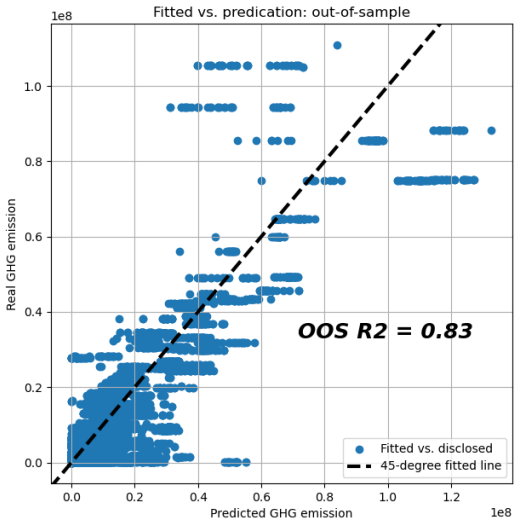


Fig. 1. Illustrative example of model training. This figure shows the partition for the training and validation set used for XGBoost learning. For similar firms that have observed carbon emission, or GHG (Greenhouse Gas) in our model, and other firm fundamentals, we use the carbon emission of disclosed 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.

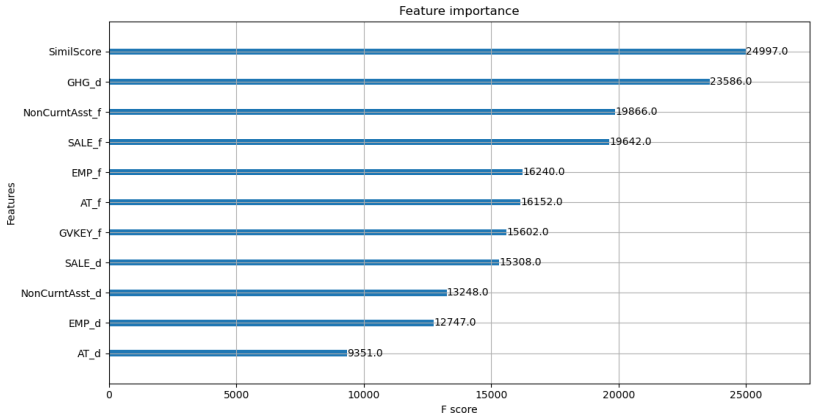


Subfigure A: XGBoost learning curve

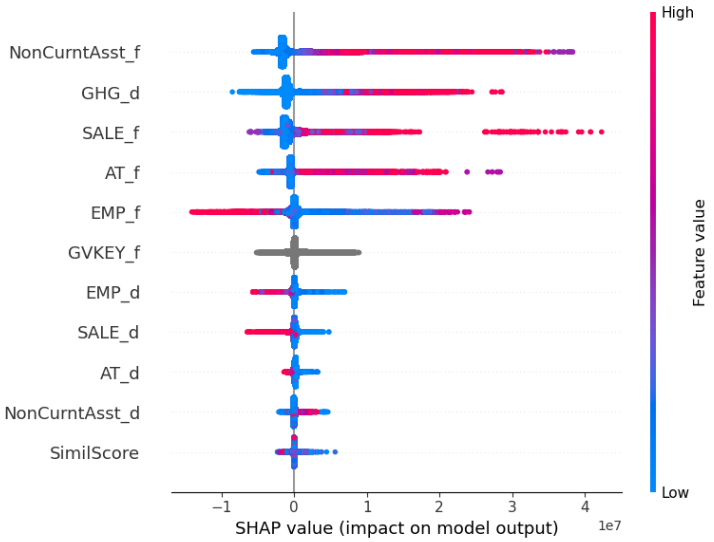


Subfigure B: Out-of-sample validation test

Fig. 2. 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. 3. 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.

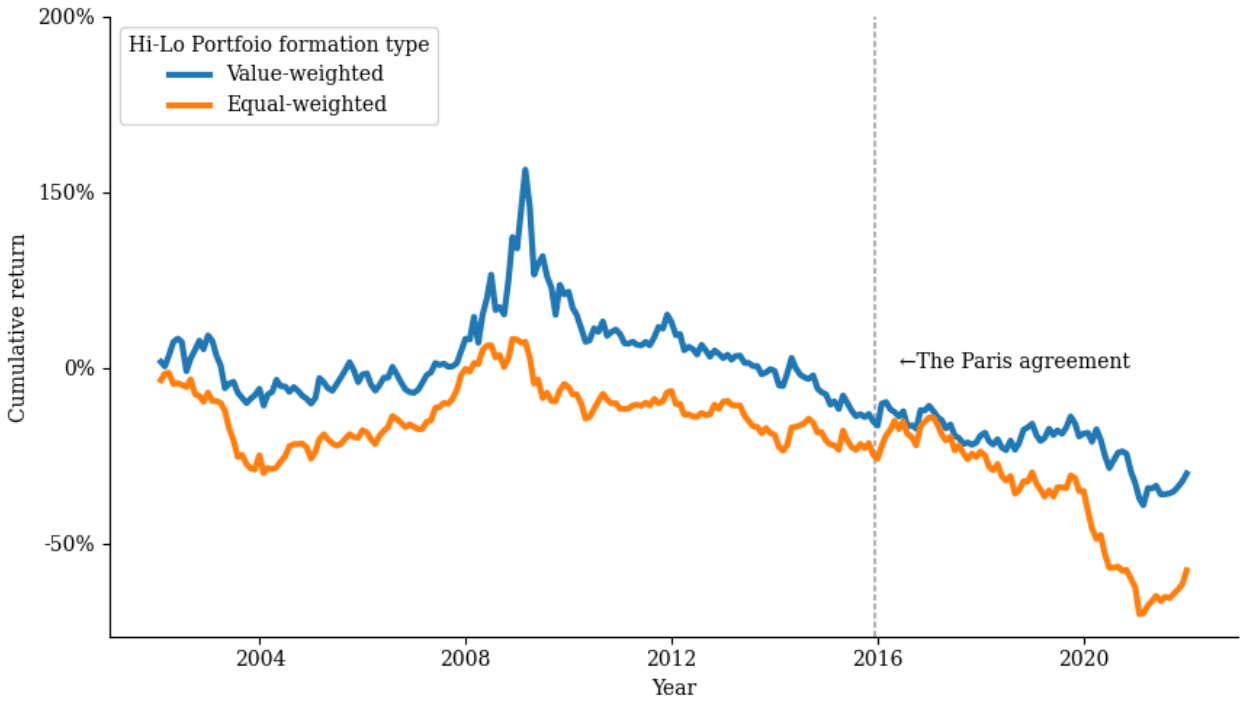


Fig. 4. Cumulative returns for high-minus-low carbon emission portfolios. This table plots cumulative returns for value-weighted or equal-weighted hi-lo portfolios sorted by logarithmic scope 1 carbon emissions at year t, where the blue line is the cumulative return of value-weighted portfolios, and the yellow line is the equal-weighted portfolio return. The time period is 2002 to 2021. The vertical dashed line denotes the announcement of the Paris agreement.

Table 1: Firm similarity score by year

	N	Mean	Std	Min	25%	50%	75%	Max
2002	86315	0.09	0.09	0.00	0.03	0.07	0.12	0.86
2003	80387	0.09	0.08	0.00	0.03	0.07	0.12	0.86
2004	78269	0.09	0.08	0.00	0.03	0.07	0.12	0.85
2005	76915	0.09	0.08	0.00	0.03	0.07	0.12	0.86
2006	76179	0.09	0.08	0.00	0.03	0.07	0.13	0.86
2007	74828	0.10	0.09	0.00	0.03	0.07	0.14	0.85
2008	70035	0.09	0.08	0.00	0.03	0.07	0.14	0.85
2009	64969	0.09	0.08	0.00	0.03	0.07	0.13	0.86
2010	62025	0.09	0.09	0.00	0.03	0.07	0.14	0.83
2011	61152	0.10	0.09	0.00	0.03	0.07	0.14	0.83
2012	60246	0.10	0.09	0.00	0.03	0.07	0.14	0.83
2013	61308	0.10	0.09	0.00	0.03	0.07	0.15	0.83
2014	64353	0.10	0.09	0.00	0.03	0.07	0.16	0.85
2015	64337	0.10	0.09	0.00	0.03	0.08	0.16	0.82
2016	62208	0.11	0.09	0.00	0.03	0.08	0.17	0.81
2017	61233	0.11	0.10	0.00	0.03	0.08	0.17	0.85
2018	61355	0.11	0.10	0.00	0.03	0.08	0.18	0.84
2019	61928	0.12	0.11	0.00	0.03	0.08	0.19	0.83
2020	61199	0.11	0.10	0.00	0.03	0.08	0.18	0.83
2021	65947	0.11	0.10	0.00	0.03	0.08	0.18	0.83

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 training and testing dataset

	N	Mean	Std	Min	25%	50%	75%	Max
$Score_{<f,d>}$	229396	0.07	0.07	0.00	0.02	0.05	0.09	0.85
LOGGHG	229396	10.59	3.33	0.00	8.43	10.40	12.65	18.92
LOGSALE	229396	7.32	2.22	0.00	6.14	7.55	8.88	13.23
LOGAT	229396	7.90	1.92	1.17	6.55	7.89	9.27	13.22
LOGNCT	229396	7.18	2.39	0.00	5.74	7.36	8.87	13.12
LOGEMP	229396	8.36	2.12	0.00	7.06	8.59	9.81	14.65

This table reports summary statistics of our dataset in three different panels. In our sample, only 229396 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. 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	9.01	5.37	11.04	32228
2003	851	11.60	2.66	11.35	9594	3703	8.96	5.09	10.80	40864
2004	1026	11.58	2.66	11.34	11699	4073	9.09	4.92	10.79	44688
2005	1260	11.37	2.68	11.19	14318	4406	9.03	4.88	10.67	46909
2006	1275	11.40	2.67	11.22	14432	4440	9.14	4.80	10.71	47311
2007	1237	11.40	2.64	11.21	13890	4367	9.24	4.78	10.82	46606
2008	1251	11.43	2.62	11.25	14025	4328	9.27	4.75	10.83	46319
2009	1265	11.30	2.62	11.07	14407	4162	9.22	4.72	10.73	46534
2010	1258	11.35	2.61	11.16	14465	4011	9.32	4.68	10.79	45738
2011	1252	11.34	2.62	11.08	14462	3938	9.35	4.67	10.80	43870
2012	1252	11.31	2.64	11.10	14484	3909	9.28	4.71	10.73	43285
2013	1350	11.25	2.65	11.07	15423	3946	9.36	4.60	10.73	43339
2014	1372	11.23	2.69	11.01	15314	4049	9.30	4.66	10.71	43275
2015	1377	11.21	2.65	11.02	15423	4117	9.36	4.51	10.68	42995
2016	3265	9.49	2.95	9.44	35521	4281	9.36	3.49	9.76	45471
2017	3286	9.48	2.98	9.41	36401	4228	9.41	3.42	9.78	45991
2018	3363	9.46	3.01	9.42	36984	4242	9.33	3.43	9.69	45611
2019	3393	9.39	3.02	9.35	36947	4279	9.29	3.39	9.62	45666
2020	3154	9.09	3.01	8.95	34567	4329	9.03	3.39	9.38	46666
2021	385	7.61	2.37	7.57	4441	4453	9.08	4.08	9.61	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

2-digit GIC name	Year-mon Obs	Mean	Std	Median	Distinct firms
Utilities	2204	14.14	2.14	14.70	168
Materials	4111	12.32	3.40	12.94	387
Energy	6073	12.16	3.43	12.58	677
Consumer Staples	2961	11.19	3.45	11.80	297
Industrials	9593	10.24	3.99	11.17	931
Consumer Discretionary	9666	9.89	3.64	10.83	1087
Communication Services	3391	9.25	3.58	9.98	379
Health Care	12312	8.93	3.62	9.83	1833
Information Technology	12768	8.52	4.11	9.92	1588
Real Estate	3240	7.10	4.46	8.86	265
Financials	15817	6.46	4.84	7.48	1737

Panel B: Detailed emission by 6-digit industry classification

6-digit GIC name	Year-mon Obs	Mean	Std	Median	Distinct firms
Multi-Utilities	452	15.14	1.28	15.77	31
Electric Utilities	819	14.96	1.53	15.81	58
Airlines	383	14.36	2.49	14.90	36
Independent Power and Renewable Electricity Producers	254	13.94	2.58	15.38	32
Construction Materials	218	13.52	2.08	13.49	18
Gas Utilities	415	13.37	1.59	13.50	29
Paper & Forest Products	264	13.02	2.21	13.33	24
Containers & Packaging	424	12.88	2.40	13.14	36
Road & Rail	716	12.43	2.63	12.48	65
Marine	231	12.42	2.57	12.89	26
Oil, Gas & Consumable Fuels	4697	12.38	3.57	12.81	529
Household Products	202	12.28	2.42	12.25	13
Industrial Conglomerates	136	12.25	4.48	14.04	11
Chemicals	1537	12.19	3.60	12.97	146
Metals & Mining	1668	12.03	3.65	12.58	163
Automobiles	230	11.86	3.31	12.64	28
Beverages	549	11.76	2.25	11.80	44
Food & Staples Retailing	552	11.50	3.29	12.00	66
Energy Equipment & Services	1376	11.40	2.81	12.00	148
Distributors	127	11.40	2.79	12.07	16
Food Products	1062	11.34	3.68	12.11	106
Water Utilities	264	11.28	2.25	11.42	18
Air Freight & Logistics	272	11.15	4.83	12.59	25
Multiline Retail	326	10.98	3.13	11.62	29
Tobacco	173	10.69	4.12	11.96	15
Construction & Engineering	601	10.67	3.13	11.36	56
Health Care Providers & Services	2033	10.43	3.49	11.32	242
Specialty Retail	2090	10.37	3.05	11.15	210
Auto Components	647	10.33	3.40	11.13	64
Aerospace & Defense	1030	10.33	3.64	11.08	99
Building Products	504	10.28	4.02	11.38	52
Hotels, Restaurants & Leisure	2295	9.99	3.51	10.88	252
Commercial Services & Supplies	1349	9.93	4.03	10.80	138
Diversified Telecommunication Services	1001	9.88	3.67	10.51	122
Semiconductors & Semiconductor Equipment	2569	9.87	3.12	10.49	243
Wireless Telecommunication Services	477	9.86	3.25	10.22	58
Machinery	1888	9.84	3.68	10.76	165
Trading Companies & Distributors	601	9.80	4.05	10.94	66
Household Durables	1078	9.78	3.96	10.84	107
Textiles, Apparel & Luxury Goods	892	9.49	3.43	10.58	92
Personal Products	423	9.34	3.75	10.31	53
Pharmaceuticals	1988	9.33	3.48	10.17	310
Diversified Consumer Services	626	9.33	3.07	10.13	75
Media	976	9.11	3.37	9.77	81
Electrical Equipment	817	9.06	3.91	10.16	94
Internet & Direct Marketing Retail	481	9.00	4.18	9.93	78
Life Sciences Tools & Services	757	8.98	3.62	10.12	87
Communications Equipment	1605	8.88	3.81	10.07	194
Technology Hardware, Storage & Peripherals	831	8.75	4.51	10.50	101
Entertainment	514	8.71	3.73	9.62	57
Health Care Equipment & Supplies	2790	8.68	3.87	9.91	357
Electronic Equipment, Instruments & Components	2230	8.61	4.48	10.14	219
Transportation Infrastructure	57	8.54	3.73	9.36	9
Biotechnology	4320	8.30	3.23	9.22	771
IT Services	1710	8.30	4.09	9.62	211
Capital Markets	2375	8.26	4.30	9.50	224
Professional Services	1008	8.23	4.32	9.63	89
Interactive Media & Services	423	8.03	3.57	8.69	61
Leisure Products	309	8.00	4.13	9.51	38
Insurance	2485	7.82	4.27	8.71	236
Health Care Technology	424	7.73	4.49	9.12	66
Software	3197	7.58	4.08	8.87	464
Diversified Financial Services	264	7.22	5.74	9.28	36
Equity Real Estate Investment Trusts (REITs)	2748	7.13	4.44	8.85	215
Real Estate Management & Development	492	6.96	4.57	9.00	50
Mortgage Real Estate Investment Trusts (REITs)	477	6.18	4.95	6.05	52
Consumer Finance	591	6.04	4.60	7.06	68
Banks	7053	5.78	4.70	6.25	705
Thrifts & Mortgage Finance	2033	5.60	5.34	5.40	289

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 6: Summary statistics and variable correlations

	N	Mean	Std	Min	25%	50%	75%	Max
Firm-yearl level observations								
LOGGHG	82213	9.22	4.44	0.00	8.03	10.43	12.02	15.87
GHGINTEN	80469	7.44	20.65	0.00	0.04	0.42	3.78	110.94
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								
RET	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

	LOGGHG	GHGINTEN	LOGSIZE	LEVERAGE	INVEST2A	ROE	HHI	LOGPPE	B2M	SALESGR	EPSGR
LOGGHG	1.00										
GHGINTEN	0.23	1.00									
LOGSIZE	0.32	-0.30	1.00								
LEVERAGE	-0.01	-0.09	0.07	1.00							
INVEST2A	0.25	0.03	0.08	-0.09	1.00						
ROE	0.08	-0.16	0.27	0.14	0.05	1.00					
HHI	0.00	0.02	-0.03	-0.08	-0.03	-0.04	1.00				
LOGPPE	0.47	-0.25	0.69	0.23	0.39	0.23	-0.05	1.00			
B2M	0.10	-0.03	-0.18	0.08	0.04	0.00	-0.01	0.23	1.00		
SALESGR	0.00	-0.01	0.08	-0.06	0.06	0.05	0.00	-0.04	-0.07	1.00	
EPSGR	0.02	-0.04	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	-0.2317***	-0.3383***		
	(-3.66)	(-5.18)		
RegulateYears			-0.0136***	-0.0211***
			(-2.93)	(-3.81)
LOGSIZE	0.3612***	0.3626***	0.3332***	0.3351***
	(10.27)	(10.25)	(7.04)	(7.06)
B2M	0.2701***	0.2728***	0.2918***	0.2941***
	(7.16)	(7.25)	(4.80)	(4.82)
LEVERAGE	1.0599***	1.0333***	0.9260***	0.9185***
	(6.28)	(6.15)	(4.18)	(4.15)
INVEST2A	-1.5413**	-1.4396**	-2.9039***	-2.9326***
	(-2.39)	(-2.24)	(-2.74)	(-2.76)
ROE	-0.0576	-0.0856	-0.0594	-0.0773
	(-0.72)	(-1.07)	(-0.52)	(-0.68)
HHI	-1.5277*	-1.5027*	-1.2401	-1.2758
	(-1.73)	(-1.71)	(-1.11)	(-1.14)
PPE	0.3098***	0.3118***	0.3061***	0.3078***
	(8.97)	(8.95)	(6.47)	(6.48)
SALESGR	-0.0128	-0.0044	0.0190	0.0268
	(-0.22)	(-0.08)	(0.25)	(0.36)
EPSGR	0.0043	0.0056	0.0008	0.0014
	(0.54)	(0.71)	(0.07)	(0.13)
Const	T	T	T	T
Ind FE	T	T	T	T
State FE		T		T
R2	0.11	0.11	0.10	0.09
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	**70.94%**	14.66%	8.78%	5.37%	1.83%
Q2 L1	13.40%	**65.16%**	17.92%	4.93%	1.09%
Q3 L1	8.37%	14.20%	**55.93%**	19.13%	2.18%
Q4 L1	5.33%	4.99%	15.69%	**60.75%**	11.56%
Q5 L1	1.95%	0.98%	1.67%	9.82%	**83.34%**
N	13952	14440	14528	14733	15157
	Panel B: Transition Prob. after 3 years				
	Q1 L0	Q2 L0	Q3 L0	Q4 L0	Q5 L0
Q1 L3	**59.10%**	19.51%	11.95%	8.34%	2.68%
Q2 L3	17.40%	**52.63%**	25.78%	8.42%	1.84%
Q3 L3	11.84%	17.91%	**41.04%**	25.33%	3.39%
Q4 L3	8.45%	7.77%	17.96%	**46.70%**	15.96%
Q5 L3	3.21%	2.18%	3.27%	11.21%	**76.13%**
N	10391	11027	11125	11676	12509
	Panel C: Transition Prob. after 5 years				
	Q1 L0	Q2 L0	Q3 L0	Q4 L0	Q5 L0
Q1 L5	**51.45%**	21.03%	14.18%	9.86%	2.91%
Q2 L5	19.99%	**46.58%**	28.76%	12.23%	2.29%
Q3 L5	14.40%	19.48%	**33.71%**	26.46%	4.68%
Q4 L5	10.15%	9.85%	19.32%	**38.50%**	18.63%
Q5 L5	4.01%	3.06%	4.04%	12.94%	**71.49%**
N	7805	8356	8500	9093	10148
	Panel D: Transition Prob. after 7 years				
	Q1 L0	Q2 L0	Q3 L0	Q4 L0	Q5 L0
Q1 L7	**46.86%**	21.33%	14.59%	10.15%	3.05%
Q2 L7	21.68%	**43.36%**	31.41%	14.25%	3.16%
Q3 L7	15.42%	20.23%	**29.55%**	27.88%	5.29%
Q4 L7	11.61%	11.20%	19.85%	**34.59%**	18.82%
Q5 L7	4.43%	3.87%	4.60%	13.14%	**69.69%**
N	5849	6248	6463	7055	8135

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	-0.0752*	-0.1465***	-0.1336***	-0.0001	-0.0027*	-0.0025*
	(-1.72)	(-3.36)	(-3.33)	(-0.08)	(-1.84)	(-1.77)
InstitOwn			0.0700***			0.0102***
			(7.38)			(6.01)
Controls	T	T	T	T	T	T
Ind 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 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		GHG_INTEN	
	(1)	(2)	(3)	(4)	(5)	(6)
LOGSIZE	0.3414***	0.3542***	-0.0038*	-0.0052	-0.3366***	-1.9057***
	(14.06)	(11.23)	(-1.86)	(-0.25)	(-5.61)	(-11.28)
B2M	0.1554***	0.177***	-0.0011	-0.0228***	-0.1300***	-0.5718***
	(12.56)	(12.47)	(-0.58)	(-3.37)	(-2.76)	(-7.30)
ROE	0.2726***	-0.0876	-0.0264***	-0.1110***	-0.1391	-5.9785***
	(5.42)	(-1.19)	(-3.52)	(-2.23)	(-1.04)	(-8.28)
LEVERAGE	0.9781***	1.0461***	0.0015	0.0233	0.1629	-6.2313***
	(8.38)	(7.08)	(0.17)	(0.41)	(0.60)	(-7.57)
INVEST2A	-4.1697***	-2.2379***	0.0464	-0.2078	-4.7228***	5.0805
	(-8.31)	(-3.49)	(0.42)	(-0.75)	(-2.36)	(1.08)
HHI	0.4701	-2.3518***	0.3405***	-0.6824*	-0.4627	2.8716
	(0.98)	(-2.38)	(3.73)	(-1.75)	(-0.56)	(0.57)
LOGPPE	0.4958***	0.3356***	0.0009	0.0007	0.3357***	-0.8784***
	(19.90)	(11.14)	(0.49)	(0.09)	(7.54)	(-5.18)
SALESGR	-0.0902*	-0.0275	0.8895***	0.5855***	-0.0828	-1.7184***
	(-1.71)	(-0.61)	(19.84)	(5.41)	(-0.64)	(-3.82)
EPSGR	-0.0044	0.0082	-0.0025**	-0.0001	0.0186	0.3005***
	(-1.07)	(0.98)	(-2.03)	(-0.02)	(1.30)	(5.59)
Const	T	T	T	T	T	T
Year FE	T	T	T	T	T	T
Ind FE	T	T	T	T	T	T
R2	0.56	0.13	0.28	0.01	0.01	0.12
N	29146	67912	26089	54992	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 firm size, 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											
Sample period	Panel A: 2002-2016						Panel B: 2002-2021						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	0.033*	0.0686***					-0.0006	0.0053					-0.0888***	-0.0536**				
	(1.89)	(3.86)					(-0.05)	(0.68)					(-3.32)	(-2.12)				
GHGGR			0.5783***	0.5721***					-0.0067	-0.0027					-0.0333	-0.0086		
			(4.84)	(5.17)					(-0.31)	(-0.13)					(-0.77)	(-0.18)		
GHGINTEN					-0.0032	-0.0039					-0.0136***	-0.0128***					-0.0368***	-0.0339***
					(-0.44)	(-0.52)					(-4.07)	(-4.54)					(-4.69)	(-5.32)
LOGSIZE	-0.0837	-0.1026	0.0189	0.0431	-0.0894	-0.0779	0.1692*	0.1514	0.2351**	0.2184**	0.138	0.1245	0.4077**	0.3759**	0.4388**	0.3837**	0.3288*	0.2828*
	(-0.75)	(-0.93)	(0.18)	(0.37)	(-0.81)	(-0.70)	(1.72)	(1.61)	(2.34)	(2.30)	(1.45)	(1.40)	(2.05)	(2.09)	(2.28)	(2.26)	(1.77)	(1.76)
B2M	-0.0480	-0.0562	-0.021	-0.0033	-0.0502	-0.0454	0.0027	0.0014	0.0022	-0.0021	-0.004	-0.0047	-0.0413	-0.0752	-0.0382	-0.0751	-0.0597	-0.1054
	(-1.31)	(-1.38)	(-0.63)	(-0.09)	(-1.37)	(-1.11)	(0.07)	(0.03)	(0.05)	(-0.05)	(-0.10)	(-0.12)	(-0.52)	(-0.95)	(-0.47)	(-0.97)	(-0.74)	(-1.32)
LEVERAGE	0.0939	-0.1837	-0.0561	0.0689	0.0641	-0.1126	0.2643	-0.0369	0.2143	-0.172	0.1872	-0.1109	0.3255	-0.9475**	0.3891	-0.8988	0.1917	-1.1771***
	(0.37)	(-0.54)	(-0.22)	(0.21)	(0.24)	(-0.33)	(0.80)	(-0.12)	(0.70)	(-0.62)	(0.56)	(-0.37)	(0.58)	(-2.28)	(0.72)	(-1.90)	(0.36)	(-2.97)
MOM	-0.0510	-0.0596	-0.0648	-0.0763	-0.0503	-0.0597	-0.32***	-0.3288***	-0.3264***	-0.3372***	-0.3179***	-0.3261***	-0.5270***	-0.5744***	-0.5205***	-0.569***	-0.5226***	-0.5675***
	(-0.53)	(-0.61)	(-0.62)	(-0.73)	(-0.53)	(-0.61)	(-3.21)	(-3.27)	(-2.89)	(-2.96)	(-3.14)	(-3.20)	(-2.90)	(-3.09)	(-2.96)	(-3.17)	(-2.73)	(-2.93)
INVEST2A	-3.3218**	-1.6153	-3.2394*	-1.8206	-3.2162**	-1.9375	-5.4709***	-3.9786***	-5.1894***	-3.3814***	-5.1657***	-3.9874***	-6.9409***	-5.2963*	-5.9519***	-3.9687	-7.3364***	-5.6558*
	(-2.25)	(-1.20)	(-1.84)	(-1.19)	(-2.15)	(-1.41)	(-3.32)	(-3.13)	(-3.0)	(-2.5)	(-3.09)	(-3.08)	(-2.97)	(-1.78)	(-2.80)	(-1.42)	(-3.33)	(-1.93)
ROE	0.8786***	0.7999***	0.6462***	0.6013***	0.8824***	0.8093***	2.1456***	1.9609***	2.1062***	1.8976***	2.0225***	1.8924***	3.0414***	2.4474***	2.9773***	2.3953**	2.8340***	2.3229***
	(3.17)	(3.56)	(2.96)	(3.06)	(3.16)	(3.59)	(9.06)	(10.97)	(8.23)	(10.31)	(8.71)	(10.61)	(7.64)	(13.56)	(7.19)	(11.80)	(6.92)	(12.53)
HHI	-0.1756	0.0888	0.7088	0.6149	-0.1974	0.1374	-1.1072	-2.529	-0.9551	-2.5869	-1.0636	-2.4719	-1.9027	-8.399	-1.5674	-6.3390	-2.2382	-7.4728
	(-0.13)	(0.06)	(0.80)	(0.40)	(-0.14)	(0.10)	(-1.30)	(-1.21)	(-0.97)	(-1.10)	(-1.26)	(-1.21)	(-0.77)	(-0.49)	(-0.60)	(-0.40)	(-0.93)	(-0.44)
LOGPPE	-0.0078	0.003	-0.0019	-0.0077	0.0307	0.0379	0.0888	0.1306***	0.0572	0.1090**	0.0825	0.1204**	0.0393	0.1489	-0.0549	0.1043	-0.0298	0.1265
	(-0.15)	(0.06)	(-0.04)	(-0.18)	(0.53)	(0.69)	(1.65)	(2.65)	(0.91)	(2.04)	(1.40)	(2.42)	(0.35)	(1.38)	(-0.42)	(0.89)	(-0.25)	(1.20)
BETA	-0.3914*	-0.4019**	-0.3483	-0.3438	-0.3953**	-0.3992**	-0.5642***	-0.5503***	-0.532**	-0.5207***	-0.5670***	-0.5455***	-0.3427	-0.2134	-0.3324	-0.2079	-0.3378	-0.2075
	(-1.96)	(-2.11)	(-1.52)	(-1.57)	(-1.98)	(-2.10)	(-2.80)	(-2.82)	(-2.63)	(-2.64)	(-2.83)	(-2.80)	(-1.21)	(-0.71)	(-1.14)	(-0.66)	(-1.21)	(-0.70)
VOLAT	0.1948***	0.2091***	0.1977***	0.2122***	0.1945***	0.2081***	0.3348***	0.3437***	0.3379***	0.3489***	0.3335***	0.342***	0.397***	0.4241***	0.3947***	0.4215***	0.3937***	0.4200***
	(3.26)	(3.46)	(2.62)	(2.78)	(3.26)	(3.44)	(6.70)	(6.96)	(6.10)	(6.33)	(6.61)	(6.87)	(5.21)	(5.52)	(5.33)	(5.64)	(4.98)	(5.31)
SALESGR	-0.2561	-0.2186	-0.2900	-0.3059	-0.2583	-0.217	-0.5819***	-0.4867***	-0.5943***	-0.4856***	-0.6134***	-0.5333	-0.7476**	-0.6291*	-0.8482***	-0.7571**	-0.7835**	-0.677*
	(-0.63)	(-0.58)	(-0.63)	(-0.68)	(-0.63)	(-0.58)	(-3.68)	(-3.04)	(-3.35)	(-2.77)	(-3.89)	(-3.38)	(-1.99)	(-1.91)	(-2.10)	(-2.13)	(-2.07)	(-2.03)
EPSGR	0.0219	0.0236	0.0111	0.0145	0.0221	0.0238	0.0777***	0.0842***	0.0431**	0.0500**	0.0828***	0.0869***	0.0298	0.0345	0.027	0.0322	0.0363	0.0369
	(0.84)	(0.85)	(0.46)	(0.59)	(0.85)	(0.87)	(3.47)	(3.76)	(2.15)	(2.56)	(3.65)	(3.84)	(0.69)	(0.87)	(0.64)	(0.85)	(0.87)	(0.98)
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	0.22	0.22	0.23	0.23	0.22	0.22	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	215808	215808	185490	185490	215760	215760	764150	764150	623506	623506	760913	760913	231149	231149	217119	217119	229777	229777

This table examines the relation between carbon emission and stock realized returns. 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 full sample period from 2002 to 2021. It has 764150 observations. The third sample period narrows down the second sample 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	-0.0093	0.0353	0.0221	0.0148	0.0008	0.0359**	0.0358	0.0491*	0.0431**	0.0225	0.025***
	(-0.34)	(1.57)	(1.24)	(1.09)	(0.06)	(2.67)	(1.65)	(1.93)	(2.52)	(1.39)	(4.47)
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	-0.0376**	-0.0004	0.0067	-0.027	-0.0925***	-0.0525**	-0.1007***	-0.0869***	-0.1807***	-0.0058	-0.0577***
	(-2.37)	(-0.03)	(0.44)	(-1.62)	(-3.63)	(-2.34)	(-4.19)	(-3.05)	(-4.22)	(-0.19)	(-3.30)
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		-0.0089	0.0354	0.0323	0.0470	-0.0001	0.0117	-0.0745	-0.0242	-0.0051	0.0015
		(-0.16)	(0.90)	(0.96)	(1.51)	(-0.00)	(0.21)	(-1.01)	(-0.59)	(-0.12)	(0.13)
R2		0.06	0.02	0.02	0.02	0.02	0.02	0.12	0.01	0.02	
N		21071	26450	31407	33184	33312	33516	33489	33490	32779	
	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	Avg coef
GHGGR	-0.0127	0.0809*	0.0819**	0.0612	-0.0349	0.0026	-0.0218	0.0197	0.1505*	-0.0661**	0.0261
	(-0.29)	(1.97)	(2.22)	(1.29)	(-0.59)	(0.06)	(-0.45)	(0.39)	(1.76)	(-2.31)	(1.31)
R2	0.01	0.03	0.01	0.01	0.03	0.01	0.02	0.03	0.12	0.10	
N	32180	31694	32210	31605	31356	36966	36935	36712	37320	37830	
Panel C: Estimated by GHGINTEN											
IndepVar.	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	Avg coef
GHG_INTEN	-0.015***	0.0008	-0.0021	-0.0014	-0.0142***	-0.0049	0.0072	-0.0121**	-0.0084**	-0.0039	-0.0055**
	(-2.78)	(0.17)	(-0.58)	(-0.48)	(-4.66)	(-1.67)	(1.46)	(-2.26)	(-2.29)	(-1.07)	(-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
GHG_INTEN	-0.0164***	-0.0174***	-0.0149***	-0.0009	-0.0341***	-0.0103**	-0.0312***	-0.029***	-0.0694***	-0.0324***	-0.0256***
	(-4.62)	(-5.11)	(-4.32)	(-0.25)	(-7.35)	(-2.45)	(-6.89)	(-5.71)	(-8.86)	(-4.85)	(-4.52)
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 and firms' carbon emissions by year following 6. 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: Emission-return relationship after the Paris agreement

	Panel A: Whole sample						Panel B: Exclude high emission industries					
Sorted by	LOGGHG		GHGGR		GHGINTEN		LOGGHG		GHGGR		GHGINTEN	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
POST*HIGHG	-0.8583*	-0.9313**	0.4629	0.4887	-1.1188***	-1.065**	-0.8922**	-0.9637**	0.4665	0.4857	-1.0997***	-1.0445**
	(-1.93)	(-2.05)	(0.89)	(0.96)	(-3.00)	(-2.51)	(-2.00)	(-2.08)	(0.90)	(0.96)	(-2.98)	(-2.43)
POST	1.0684***	1.1492***	4.1305***	4.251***	0.5041	0.5048	1.1873***	1.237***	4.123***	4.2275***	0.4106	0.3976
	(3.38)	(3.25)	(9.21)	(8.46)	(1.52)	(1.29)	(3.65)	(3.44)	(8.87)	(8.21)	(1.23)	(1.01)
HIGHG	0.3554*	0.4209***	0.2523**	0.2596**	-0.0238	-0.0064	0.3643*	0.4503***	0.2457**	0.2604**	-0.0163	0.0156
	(1.75)	(3.44)	(2.36)	(2.30)	(-0.18)	(-0.06)	(1.81)	(3.69)	(2.36)	(2.37)	(-0.12)	(0.14)
Const	T	T	T	T	T	T	T	T	T	T	T	T
Controls	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	0.16	0.16	0.14	0.15	0.16	0.16	0.16	0.16	0.14	0.15
N	308838	308838	241463	241463	295656	295656	300021	300021	233743	233743	287194	287194

This table examines the emission-return relationship after the Paris agreement. We first sort firms by their carbon emissions with three different measures into five quintiles from low to high. We keep the lowest and the highest quintile and define a dummy variable that indicates whether a firm is a high-emission firm if it belongs to the high-emission quintile. We also define a time dummy that indicates whether after the Paris agreement. The interaction term between the two dummies is of interest. Similar to 6, the dependent variable is stock return, and other financial variables 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-fixed effects and industry-fixed effects and cluster standard errors at the 2-digit GIC industry and year levels in the regressions. We use the full sample estimated by the XGBoost model in the left panel and use the sample excluding salient industries (Oil, Gas & Consumable Fuels, Gas Utilities, and Transportation) in the right panel.

Table 14: The low-carbon premium and common risk factors

	Panel A: 2002-2021						Panel B: 2016-2021 (after the Paris agreement)					
Estimated by	LOGGHG		GHGGR		GHGINTEN		LOGGHG		GHGGR		GHGINTEN	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Intercept	-0.0209	-0.0135	0.0214*	0.0158	-0.012***	-0.0132***	-0.0957***	-0.0796***	0.0274	0.0171	-0.0272***	-0.0288***
	(-1.53)	(-1.06)	(1.79)	(1.36)	(-4.01)	(-4.35)	(-8.02)	(-4.85)	(1.34)	(0.71)	(-5.68)	(-6.23)
RMRF	0.0013		-0.0003		-0.0006		0.0093***		-0.0151***		-0.0001	
	(0.64)		(-0.10)		(-1.17)		(2.95)		(-2.81)		(-0.10)	
SMB	-0.0079		-0.0022		-0.0037***		-0.0218***		0.0032		-0.0081***	
	(-1.54)		(-0.33)		(-3.49)		(-2.56)		(0.27)		(-6.05)	
HML	-0.0065		-0.0146**		-0.0004		-0.0036		-0.0224***		0.0004	
	(-1.42)		(-2.29)		(-0.30)		(-0.53)		(-3.13)		(0.23)	
RMW	0.0055		-0.0064		0.0003		0.0082		0.0131		-0.0023	
	(1.16)		(-1.12)		(0.23)		(1.20)		(1.00)		(-1.12)	
CMA	0.0153**		-0.0146*		-0.0022*		0.0256**		-0.0166*		-0.0034**	
	(2.57)		(-1.69)		(-1.80)		(2.37)		(-1.96)		(-2.10)	
BAB	0.0046		-0.0057		0.0001		0.0052		-0.0119		-0.0003	
	(1.31)		(-1.23)		(0.01)		(0.76)		(-1.10)		(-0.17)	
LIQ	0.001		-0.0032		0.0003		-0.0035		0.0116		-0.0011**	
	(0.39)		(-0.86)		(0.53)		(-0.95)		(1.66)		(-1.93)	
Mom	0.0001		0.0008		-0.0001		0.0046		-0.0042		-0.0020*	
	(0.05)		(0.29)		(-0.27)		(1.10)		(-0.47)		(-1.89)	
R2	0.10	0.00	0.10	0.00	0.14	0.00	0.33	0.00	0.28	0.00	0.33	0.00
N	240	240	228	228	240	240	72	72	72	72	72	72

This table examines the low-carbon return premium after controlling for common risk factors. The dependent variable is the monthly low-carbon premium estimated using a cross-sectional return regression in 6. 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 premiums on common risk factors, including the market factor as the CAPM model in panel A. 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. Similar business structures and similar carbon emission

Despite the intuitiveness, it still remains to be investigated whether firms with similar business structures share similar emission patterns. There may exist industry-specific heterogeneity that prevents the algorithm to make 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 design 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 0.83 in the main analysis (see subfigure B in figure 2) to 0.70 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 remain 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 0.19. We do not observe a very lower R-square is because we inlcude both the firm fixed effects and firm fundamentals into the algorithm.

[Insert Figure A1 near here]

A.2. Different data partition 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 basically 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 time 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), they compare 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 time 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 A6.

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.

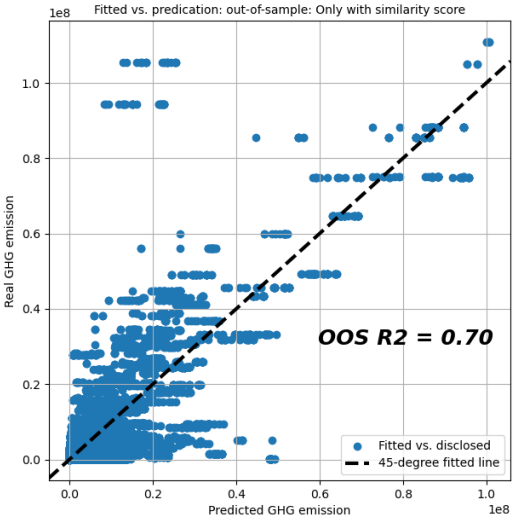
[Insert Figure A5 near here]

A.5. Emission autocorrelation patterns

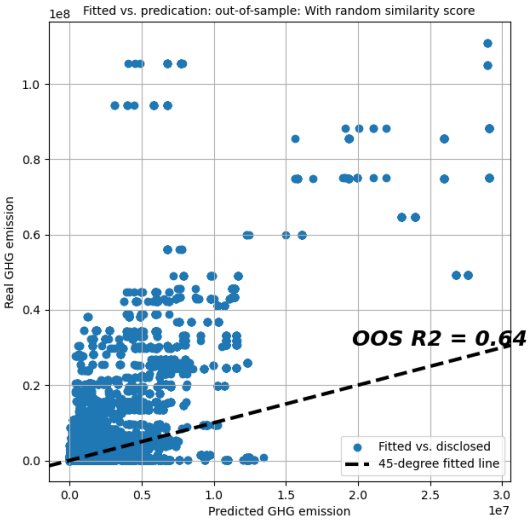
We show that carbon emission is quite persistent with data sample estimated by the XGBoost algorithm in table 8, as firms are more likely to stay within the emission quintiles 1/3/5/7 years after the formation date. We also perform auto-correlation regressions to examine the persistence of carbon emissions. 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 0.6881 and 0.5448 without and with controls, and t-statistics are 35.21 and 24.15, 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]

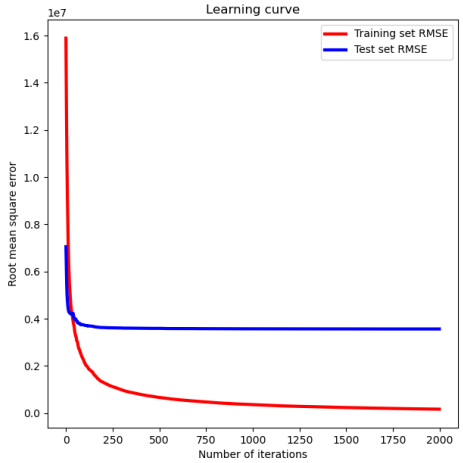


Subfigure A: Prediction only with similarity score

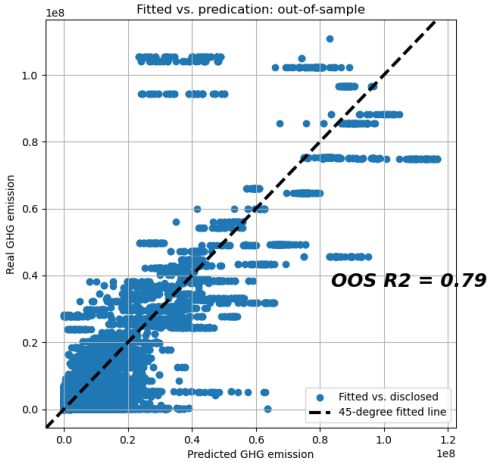


Subfigure B: Prediction with random similarity score

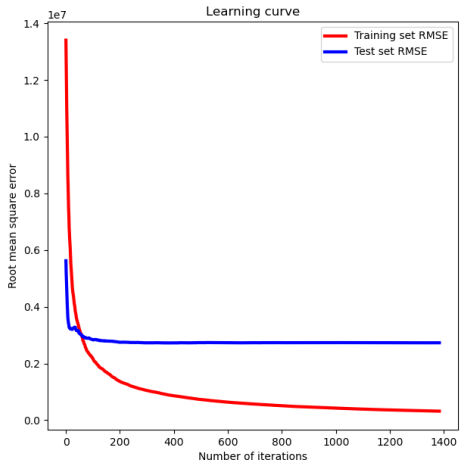
Fig. A1. Business similarity and emission similarity. In subfigure A, we only include firms' business similarity scores and their fixed effects (GVKEY) in both the training set and the test set. We use the same training method as well as machine learning parameters as in the main analyses and rerun the 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.



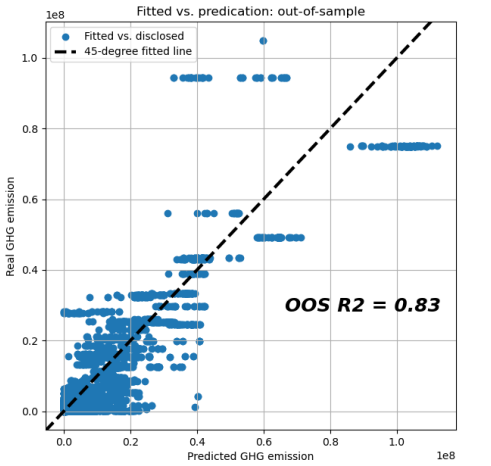
(a) Learning curve (Training set: 2002-2017)



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

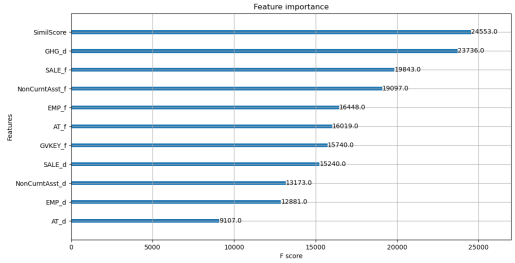


(c) Learning curve (Training set: 2002-2019)

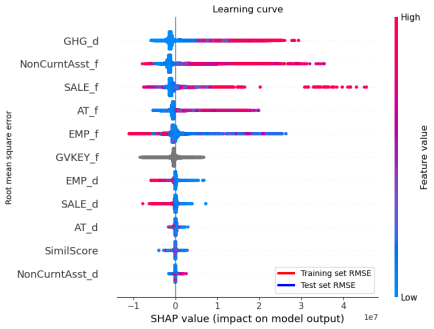


(d) OOS validation (Training set: 2002-2019)

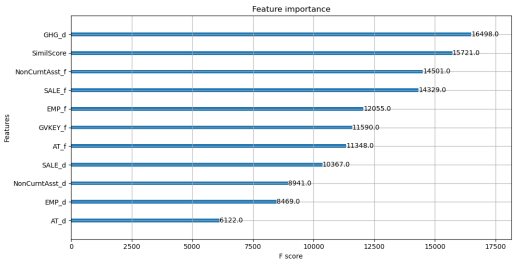
Fig. A2. XGBoost performance results with different training periods. In subfigures (a) and (b), we set the training set from 2002 to 2017, and the test set from period 2018 to 2021. In subfigures (c) and (d), we choose the training set from 2002 to 2019, and the test set from the period 2020 to 2021. 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.



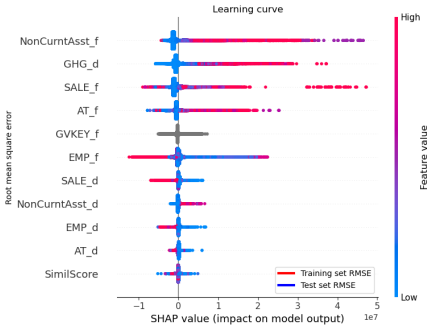
(a) Importance plot (Training set: 2002-2017)



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

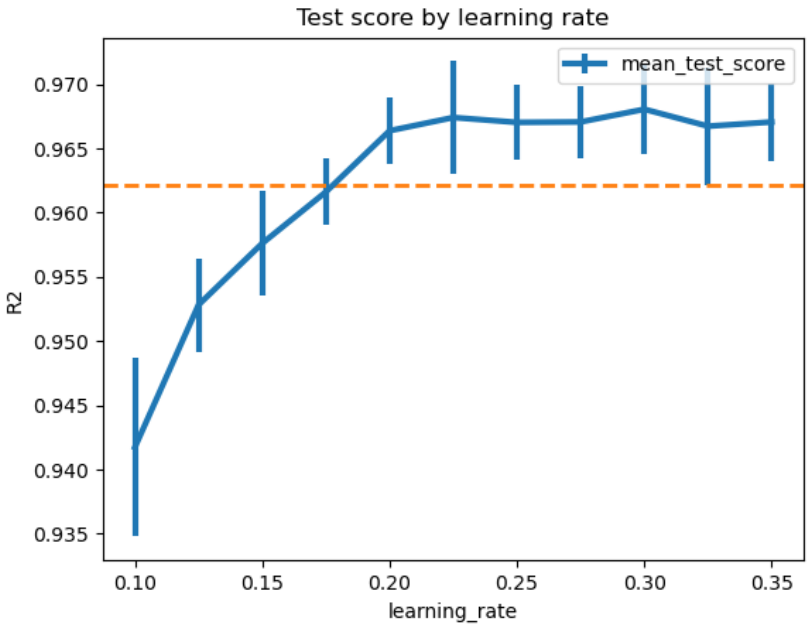


(c) Importance plot (Training set: 2002-2019)

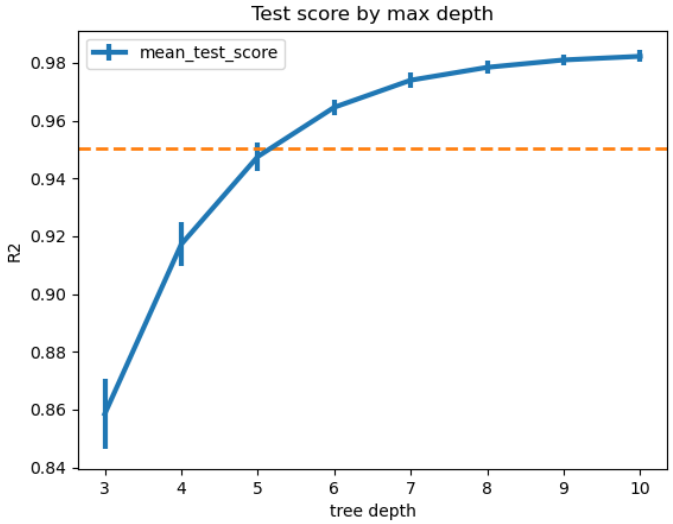


(d) SHAP plot (Training set: 2002-2019)

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 subfigure (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 subfigure (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.

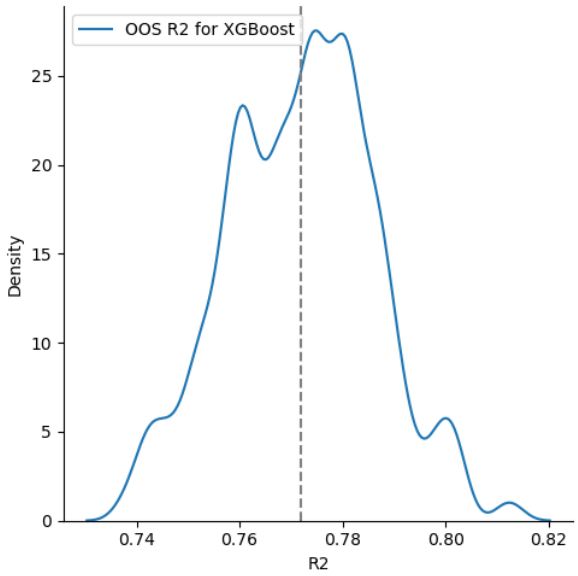


Subfigure A: Cross-validation test on learning rate

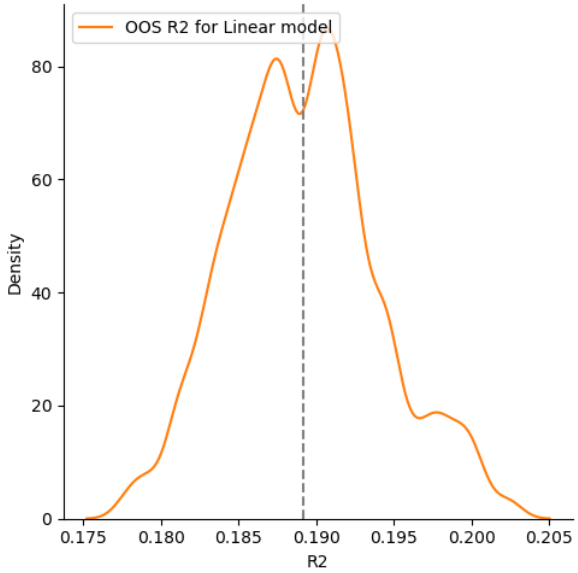


Subfigure B: Cross-validation test on tree depth

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. 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

	$LOGGHG_t$		$GHGGR_t$		$GHGINTEN_t$	
	(1)	(2)	(3)	(4)	(5)	(6)
$LOGGHG_{t-1}$	0.6881***	0.5448***				
	(35.21)	(24.15)				
$GHGGR_{t-1}$			-0.0974***	-0.1088***		
			(-13.87)	(-15.96)		
$GHGINTEN_{t-1}$					0.7407***	0.7024***
					(40.16)	(33.66)
Const	T	T	T	T	T	T
Control		T		T		T
Year FE	T	T	T	T	T	T
R2	0.49	0.40	0.01	0.02	0.59	0.58
N	76113	63463	54499	48103	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 empirical results

B.1. Low-carbon premium with Trucost data after 2016

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 dataset 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 not reliable and unconvincing. In table B1, we estimate the emission-return relationship with full sample data from 2002 to 2021 as well as the post-Paris agreement sample period provided by the Trucost database.

Regression results show that, in Panel A, where we use the whole sample period, 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. In panel B where we restrict the sample period from 2017 to 2021, the coefficient of interest become significantly negative with -0.0898 and t-statistic -3.66 in column 7. Similar to table 11, the negative emission-return relationship becomes insignificant once we include the industry-fixed effects in column 8.

In columns 3 to 4 and 9 to 10 where the independent variable is the emission growth rate, the positive relationship remains stable post the Paris agreement. As we show in the next section with sorting results (in table B2), 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. 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.

Overall, the main message we would like to convey is that the carbon premium might be no longer significant, if not totally reversed, 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 might be expected to observe different results in the next coming years.

[Insert Table B1 near here]

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 B2. Sorting results suggest that on average high-emission stocks underperform low-emission stocks by 0.09% and 0.3% 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. We also plot cumulative portfolio returns similar to figure 4 with the Trucost data in figure B1. This figure also shows that the negative emission-return relationship is more pronounced after the Paris agreement.

[Insert Figure B1 near here]

[Insert Table B2 near here]

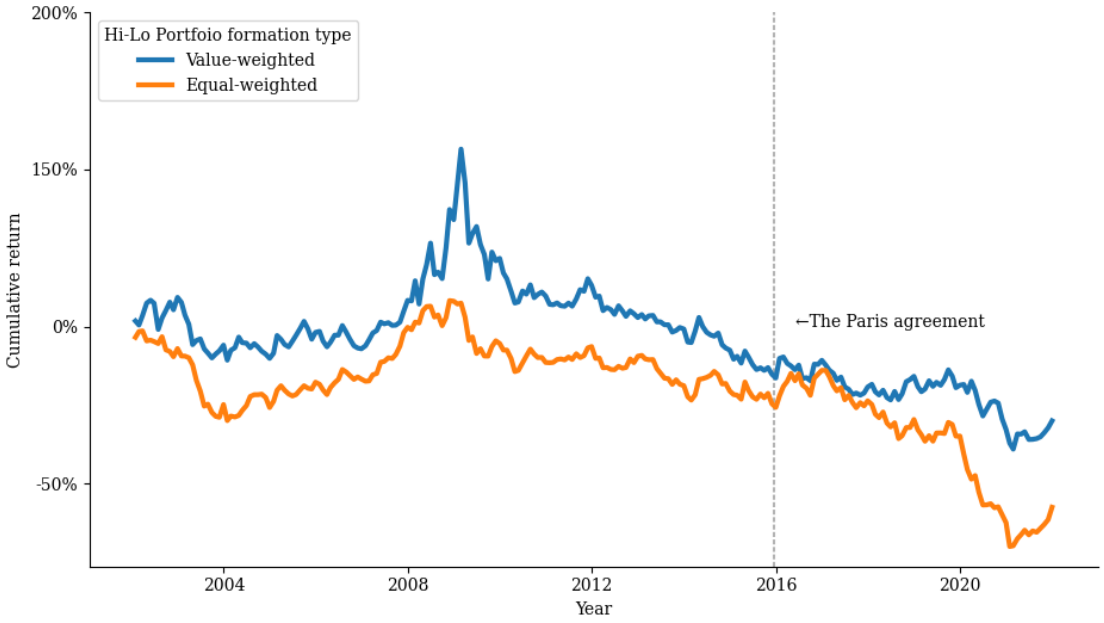


Fig. B1. Cumulative returns for high-minus-low carbon emission portfolios estimated with Trucost data. This table plots cumulative returns for value-weighted or equal-weighted hi-lo portfolios sorted by logarithmic scope 1 carbon emissions at year t, where the blue line is the cumulative return of value-weighted portfolios, and the yellow line is the equal-weighted portfolio return. The time period is 2002 to 2021. The vertical dashed line denotes the announcement of the Paris agreement.

Table B1: Carbon emission and realized stock returns with data only from Trucost

	Trucost data sample											
Sample period	Panel A: 2002-2021						Panel B: 2016-2021 (after the Paris agreement)					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
LOGGHG	0.0098	0.0536**					-0.0898***	-0.0085				
	(0.48)	(2.34)					(-3.66)	(-0.16)				
GHGGR			0.7061***	0.7412***					0.9119*	1.0895**		
			(3.72)	(3.81)					(1.87)	(2.09)		
GHGINTEN					-0.0081	-0.0127					-0.0423	-0.0332
					(-0.76)	(-1.00)					(-0.86)	(-0.84)
LOGSIZE	0.2110	0.1799	0.3317**	0.3280**	0.2034	0.1959	0.7525***	0.6214**	0.7759***	0.6725**	0.7164***	0.6085***
	(1.31)	(1.15)	(2.06)	(2.03)	(1.30)	(1.30)	(3.29)	(2.48)	(3.26)	(2.77)	(3.44)	(2.73)
B2M	0.0051	-0.0071	0.0356	0.0401	0.0026	0.0004	0.0373	0.0056	0.0495*	0.026	0.0305	0.0014
	(0.15)	(-0.20)	(1.13)	(1.29)	(0.08)	(0.01)	(1.34)	(0.18)	(1.84)	(0.88)	(1.07)	(0.05)
LEVERAGE	0.0248	-0.3707	-0.0686	-0.2306	0.0132	-0.3141	-0.2922	-1.1791***	-0.0177	-1.0527***	-0.1547	-1.1884***
	(0.11)	(-1.40)	(-0.30)	(-0.86)	(0.06)	(-1.20)	(-0.59)	(-3.02)	(-0.05)	(-3.42)	(-0.30)	(-3.03)
MOM	-0.1588	-0.1777	-0.1819	-0.2036	-0.1584	-0.1775	-0.3415	-0.403	-0.3415	-0.4052	-0.3409	-0.4025
	(-1.47)	(-1.56)	(-1.56)	(-1.65)	(-1.47)	(-1.56)	(-1.54)	(-1.60)	(-1.53)	(-1.60)	(-1.52)	(-1.59)
INVEST2A	-4.1423**	-1.1284	-4.2961***	-1.3298	-4.0824**	-1.4080	-4.0486*	-2.1562	-5.2882*	-2.9505	-4.2593*	-2.1882**
	(-2.66)	(-0.94)	(-2.47)	(-1.13)	(-2.65)	(-1.17)	(-1.76)	(-1.02)	(-2.01)	(-1.44)	(-1.97)	(-1.10)
ROE	1.3541***	1.2599***	1.3136***	1.2813***	1.3487***	1.2730***	1.9664***	1.7233***	2.0597***	1.8355***	1.9064***	1.7241***
	(4.88)	(5.24)	(4.58)	(4.81)	(4.95)	(5.30)	(5.84)	(5.06)	(5.89)	(6.19)	(5.75)	(5.23)
HHI	0.0808	-1.3239	0.5122	-1.4447	0.0584	-1.2548	0.2043	-5.9436	-0.2678	-8.5278	-0.0536	-5.8526
	(0.07)	(-0.77)	(0.54)	(-0.85)	(0.05)	(-0.73)	(0.09)	(-1.33)	(-0.12)	(-1.91)	(-0.02)	(-1.28)
LOGPPE	-0.0698	-0.0277	-0.0822	-0.0303	-0.0500	0.0026	-0.1871*	-0.0311	-0.2564**	-0.0378	-0.2349***	-0.0253
	(-1.30)	(-0.60)	(-1.41)	(-0.62)	(-0.88)	(0.06)	(-1.89)	(-0.40)	(-2.67)	(-0.42)	(-3.35)	(-0.29)
BETA	-0.5782**	-0.5715**	-0.5272**	-0.5147**	-0.5815**	-0.5690	-0.5881*	-0.4872	-0.5841*	-0.4947	-0.5983*	-0.4839
	(-2.66)	(-2.73)	(-2.30)	(-2.35)	(-2.65)	(-2.74)	(-1.76)	(-1.54)	(-1.76)	(-1.58)	(-1.75)	(-1.54)
VOLAT	0.3311***	0.3514***	0.3333***	0.3532***	0.3310***	0.3510***	0.4642***	0.5014***	0.4479***	0.4873***	0.4625***	0.5011***
	(3.49)	(3.64)	(3.37)	(3.51)	(3.49)	(3.64)	(3.41)	(3.51)	(3.19)	(3.31)	(3.37)	(3.51)
SALESGR	-0.5847*	-0.4992*	-0.6767**	-0.6188**	-0.5850**	-0.4968*	-1.0494**	-0.8052	-1.1018**	-0.8971*	-1.0314**	-0.8129
	(-2.04)	(-1.82)	(-2.23)	(-2.14)	(-2.05)	(-1.81)	(-2.28)	(-1.52)	(-2.71)	(-2.02)	(-2.21)	(-1.55)
EPSGR	0.0510*	0.0525*	0.0375	0.0379	0.0513*	0.0523*	0.0694	0.0771	0.0552	0.0626	0.072	0.077
	(1.75)	(1.71)	(1.33)	(1.30)	(1.75)	(1.72)	(1.59)	(1.42)	(1.24)	(1.16)	(1.55)	(1.43)
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	0.23	0.23	0.23	0.24	0.23	0.23
N	332410	332410	299871	299871	332338	332338	116602	116602	114381	114381	116578	116578

This table examines the relation between carbon emission and stock realized returns with full sample data provided by Trucost. 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 data period 2002 to 2021. In panel B, we use a subsample that starts after the Paris agreement from 2017 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 B2: 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	0.6336*	1.1785***	3.98	-0.16	1.39	746.49	1980.58	520.75	12.55	0.49	0.73	0.03	-0.08	0.10	3.18	0.12	-0.13
	(1.78)	(2.81)															
2	0.8837***	1.1544***	9.74	0.30	6.59	1515.48	2453.22	1303.02	13.11	0.50	0.88	0.04	-0.04	0.10	4.04	0.11	-0.12
	(2.88)	(2.85)															
3	0.7291**	1.0347**	10.98	0.51	9.03	2322.34	3121.30	1957.12	13.35	0.54	0.81	0.05	0.00	0.10	4.75	0.11	-0.08
	(2.31)	(2.51)															
4	0.7652***	1.0837***	12.20	0.62	9.97	5541.45	7575.81	5362.59	13.88	0.58	1.06	0.06	0.07	0.10	5.92	0.10	0.01
	(2.92)	(2.80)															
Hi	0.5343**	0.8782**	14.55	0.54	11.21	19320.72	27068.58	19925.85	14.99	0.64	1.69	0.06	0.10	0.09	7.87	0.08	0.07
	(2.02)	(2.36)															
Hi-Lo	-0.0993	-0.3003															
	(-0.49)	(-1.41)															

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	1.0090***	1.4230***	7.62	0.13	0.14	1745.68	3190.49	1785.24	13.74	0.52	0.92	0.03	-0.06	0.11	4.16	0.15	-0.06
	(2.90)	(3.55)															
2	0.9086***	1.2222***	9.82	0.07	0.23	3236.39	5277.20	3343.84	14.46	0.55	1.02	0.04	0.08	0.10	5.75	0.09	0.01
	(2.92)	(3.32)															
3	0.8106***	1.1244***	11.10	0.06	0.47	7270.52	9679.34	6797.38	14.89	0.59	1.11	0.05	0.11	0.10	6.57	0.08	0.02
	(2.84)	(2.98)															
4	0.6015**	0.8902**	12.57	0.08	1.58	15009.60	20032.51	13627.19	15.17	0.62	1.50	0.06	0.12	0.10	7.58	0.07	0.03
	(2.38)	(2.52)															
Hi	0.6252**	0.9052**	15.42	0.10	10.56	27235.55	38492.71	29081.39	15.40	0.65	2.35	0.07	0.09	0.09	8.67	0.06	0.05
	(2.22)	(2.38)															
Hi-Lo	-0.3837*	-0.5178**															
	(-1.69)	(-2.25)															

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.