

From Stocks to Sustainability: Predicting Carbon Emissions with Machine Learning Models

Divya Chaudhary, Saanidhya Vats, Anjali Haryani, Siva Sai Gopaal Praturi

Northeastern University, Seattle, WA, USA

d.chaudhary@northeastern.edu, vats.saa@northeastern.edu, haryani.a@northeastern.edu, praturi.s@northeastern.edu

Abstract

Global warming and rising carbon emissions pose serious environmental challenges, necessitating the development of new solutions to reduce their effects. As financial markets become more environmentally conscious, knowing the relationship between stock prices and carbon emissions has become critical for making long-term decisions. This paper introduces a novel dataset, finance and emission data, capturing the relationship between companies' stock values and their corresponding yearly carbon emissions. Using finance and emission data, the study aims to forecast carbon emissions based on stock value trends using five machine learning models: DistilBERT, Long Short-Term Memory (LSTM), Convolutional Neural Network-LSTM (CNN-LSTM), Prophet and DistilBERT-LSTM. The models are trained and validated on the finance and emission data, and their performance is compared using four key metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). Among the models evaluated, DistilBERT outperformed others, achieving an average MAE of 2.7960, MSE of 11.7369, RMSE of 3.4271, and MAPE of 24.2709. These results demonstrate the potential of utilizing stock market data for predictive environmental analysis and highlights the effectiveness of transformer-based models for time-series forecasting tasks.

Introduction

In recent years, the growing urgency of climate change has emphasized the need for innovative strategies to monitor and reduce carbon dioxide (CO₂) emissions (IPCC 2018; United Nations 2020). As a major contributor to global warming, CO₂ emissions must be accurately forecasted to guide policymakers, businesses, and environmental organizations in achieving sustainability targets and adhering to international climate agreements (Stern 2007; International Energy Agency 2021). Traditional emission prediction methods often rely on historical industrial output and energy consumption data but may overlook broader economic indicators, such as stock price fluctuations, which can reflect underlying economic activities (Zhang et al. 2019; Chen and Wang 2020).

The interplay between financial markets and environmental impact has gained attention in academic and industry circles (Porter and Kramer 2011; Eccles, Ioannou, and Serafeim 2014). Stock prices, as proxies for corporate performance, capture investor sentiment and expectations, including environmental practices (Friede, Busch, and Bassen 2015; Eccles and Serafeim 2013). This study explores the potential of stock prices to predict CO₂ emissions by developing a time series model leveraging historical stock price data.

Time series analysis is uniquely suited for identifying temporal patterns and trends in stock prices and emission levels (Box and Jenkins 1976; Hyndman and Athanasopoulos 2018). By employing advanced deep learning techniques, renowned for their sequential data modeling capabilities (Hochreiter and Schmidhuber 1997; Graves 2012), this research aims to improve the accuracy of emission forecasts.

The primary objective is to establish a robust predictive model correlating stock price movements with carbon emissions while evaluating the effectiveness of various time series approaches. By integrating financial indicators into environmental forecasting, this study seeks to provide actionable insights for investors, policymakers, and corporate leaders (Goodfellow, Bengio, and Courville 2016; Chollet 2017).

Ultimately, this research bridges the fields of finance and environmental science, offering a novel approach to understanding and predicting carbon emissions. The findings have the potential to inform sustainability strategies and highlight the interconnectedness of economic activities and environmental impact (World Economic Forum 2022; McKinsey Global Institute 2020).

This research addresses the challenge of predicting stock prices based on corporate GHG emissions using advanced machine learning models: DistilBERT, Long Short-Term Memory (LSTM), CNN-LSTM, Prophet and DistilBERT-LSTM. Performance is evaluated using four metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE).

The study is divided into various sections: Related Work examines earlier studies and contextualizes the contribution of this research. Data Creation describes the integration of NASDAQ financial and GHGP emissions data. Data preprocessing explains the cleaning, matching, and standardizing

procedures used. Data Transformation describes the structure of longitudinal data for trend analysis. Model Selection describes the chosen model architectures, training procedures, and evaluation frameworks, while Results provides a comparative study of model performance and major conclusion.

This structured approach ensures a comprehensive analysis, supporting data-driven decision-making at the nexus of environmental and financial performance.

Related Work

In the time of increased environmental awareness, the financial landscape is experiencing a significant shift. Investors are recognizing the intricate link between a company's environmental impact and its market performance, with greenhouse gas (GHG) emissions emerging as a critical factor in stock market dynamics. This integration of environmental metrics with financial analysis is reshaping investment strategies and corporate decision-making, offering actionable insights for stakeholders across the board.

Research works have progressively shed light on this connection. (Tran et al. 2024) demonstrates elevated stock price crash risks for carbon-intensive firms, particularly under democratic presidencies. While, their findings highlight the financial implications of carbon risk, they do not incorporate predictive modeling. Our study addresses this gap by developing machine learning frameworks that forecast stock prices using emissions data across various industries.

Other research, such as (Salehi et al. 2022), investigates the impact on Iranian automotive firms, revealing an inverse relationship between emissions and financial performance. However, the industry-specific focus limits broader applicability. In contrast, our approach extends this work by applying advanced deep learning techniques to NASDAQ-listed firms, offering wider relevance.

The dual role of financial development and renewable energy consumption is examined by (Paramati, Mo, and Gupta 2017), who find that stock market growth drives emissions reductions in developed economies. Similarly, (Zafar et al. 2019) highlight the divergent effects of market development on emissions across G-7 and N-11 countries. While both studies underscore emissions mitigation, they do not establish connections to stock price predictions. Our research bridges this gap by employing machine learning to integrate emissions trends with financial outcomes.

The concept of emissions trading schemes adds another layer of insight. (Wen, Wu, and Gong 2020) reveal a "carbon premium," where firms involved in emissions trading experience higher stock returns. (Oestreich and Tsiakas 2015) report similar findings in German companies under the EU ETS, attributing increased returns to free carbon allowances. While these studies provide valuable insights, their retrospective nature does not support predictive modeling. Our work builds upon this research by using models to project stock price trends based on emissions data.

Policy impacts have also been explored. (Hengge, Panizza, and Varghese 2023) analyze carbon policy effects in Europe, noting that stricter regulations increase costs for

emission-heavy firms. This complements studies like (Tang and Li 2022), which illustrate the potential of Multi-CNN models in forecasting stock returns for carbon-intensive industries. However, the latter does not account for different sectors or time-series dynamics. Our research broadens the scope by integrating diverse industries and dynamic time-series models for comprehensive forecasting.

Recent work has delved into sector-specific insights. (Aswani, Raghunandan, and Rajgopal 2023) analyze emissions intensity among U.S. firms, linking it to company size rather than efficiency. Meanwhile, (Chang et al. 2020) identify bidirectional causality between CO2 emissions and stock returns, focusing on financial impacts but not predictive capabilities. By directly incorporating emissions data, our study uncovers actionable trends in stock price movements.

(Antoniuk and Leirvik 2024) analyze the impact of carbon prices on corporate GHG emissions across 1591 firms in 23 European countries. They find that higher carbon prices lead to a reduction in emissions, with a more pronounced negative effect observed during Phase 3 of the EU ETS. Although this work highlights carbon pricing's impact on emissions, it does not engage in predictive modeling for financial markets. Our research extends these insights by using machine learning models to integrate carbon price impacts with stock price forecasts.

(Giovanna Bua and Rognone 2024) develop novel indicators for physical and transition climate risks, showing their economic significance post-2015. They observe that increases in transition risks are often linked to higher returns for green stocks, while physical risks typically lead to lower returns for brown stocks. However, their study does not apply these risk indicators to predictive financial models. Our approach builds on their findings by leveraging machine learning techniques to forecast stock trends informed by climate risk data.

(Adamolekun 2024) utilize event study methodology to investigate how unexpected political events impact climate-sensitive sectors. They find that events such as the Paris Agreement and Climategate positively influenced the clean energy sector, while events undermining climate policies benefited the fossil energy sector. While this research shows how political events affect market behavior, it does not integrate these insights into predictive stock models. Our approach enhances this by including climate policy impacts, improving the predictive power of financial models.

This comprehensive narrative underscores a transformative shift: the link between GHG emissions and stock market behavior is becoming increasingly significant. By leveraging machine learning methodologies our study enhances the understanding of this relationship.

Finance and Emission Data

Data Creation

This research integrates two primary datasets to examine the relationship between environmental emissions and stock performance among publicly traded companies in the U.S. By linking facility-level greenhouse gas emissions data with

company-level financial data, a multi-dimensional analysis of corporate environmental impact and financial outcomes over time was achieved.

NASDAQ Company Screener Dataset: This dataset, provided in CSV format, includes comprehensive information on companies listed on the NASDAQ stock exchange. Each record represents a publicly traded company, with key fields such as:

- **Symbol:** The stock ticker symbol of the company, used as a unique identifier.
- **Name:** The full name of the company.
- **Additional Fields:** Attributes such as sector and industry classification that contextualise the company's financial and operational background.

Greenhouse Gas Protocol (GHGP) Facility-Level Data: This dataset, sourced from the Greenhouse Gas Protocol, reports annual greenhouse gas emissions for individual facilities across various industries from 2011 to 2018. It provides a longitudinal view of facility-level emissions, enabling year-over-year trend analysis. Key fields include:

- **Facility Name:** The official name of the facility reporting emissions.
- **Reported Emissions:** The total greenhouse gas emissions reported by each facility for each year.
- **Location and Sector Information:** Fields indicating the facility's location (state, county) and industrial sector, which provide context for emissions data and facilitate sector-level comparisons.

Data Preprocessing

To prepare the Finance and Emission Data for integration, the following preprocessing steps were performed:

- **Name Cleaning and Standardization:** Variations in names due to legal suffixes (e.g., "Inc.", "Corp.") and special characters were removed. Specific steps included:
 - Lowercasing all names for case consistency.
 - Removing punctuation and non-alphanumeric characters.
 - Eliminating common legal suffixes (e.g., "Inc.", "Corp.") using regular expressions.
- **Vectorization with TF-IDF:** Cleaned names were vectorized using Term Frequency-Inverse Document Frequency (TF-IDF) to emphasise distinctive terms, reducing noise and improving matching accuracy.
- **Cosine Similarity Calculation:** Cosine similarity was computed between NASDAQ company names and GHGP facility names to evaluate alignment. A similarity threshold of 0.78 was applied to retain high-probability matches.

Merged Finance and Emission Data Characteristics

The merged Finance and Emission Data includes the following fields:

- **Symbol:** The NASDAQ ticker symbol of the matched company.
- **Company Name:** The cleaned and standardised name of the NASDAQ-listed company.
- **Matched Facility:** The facility name from the GHGP dataset that achieved a similarity score above the threshold.
- **Similarity Score:** The cosine similarity score between the company and facility names, indicating the strength of the match.

This integration was repeated for each year of GHGP emissions data (2011–2018), creating a comprehensive longitudinal Finance and Emission Data. Figure 1 shows the emissions over a time period of 8 years (2011 to 2018).

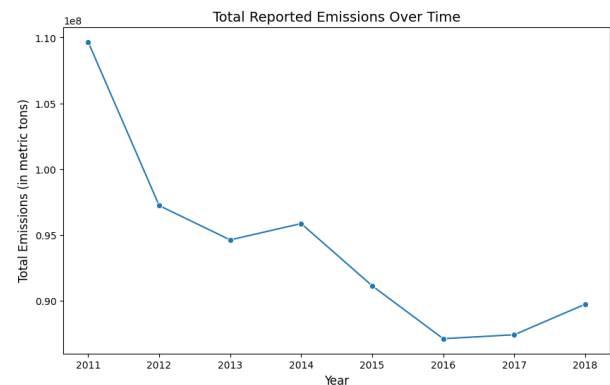


Figure 1: Emissions over time from 2011 to 2018.

Financial Data Integration

For each matched company-year pair, annual average stock prices were calculated from historical daily prices obtained from Yahoo Finance. Metrics included:

- **Average High Price:** The mean of the daily high prices over the year.
- **Average Low Price:** The mean of the daily low prices.
- **Average Open Price:** The mean of the daily opening prices.
- **Average Close Price:** The mean of the daily closing prices.

Stock prices were added to the corresponding emissions data, enriching the Finance and Emission Data with financial metrics for detailed analyses. The Figure 2 illustrates a fine comparison between emissions and stock prices.

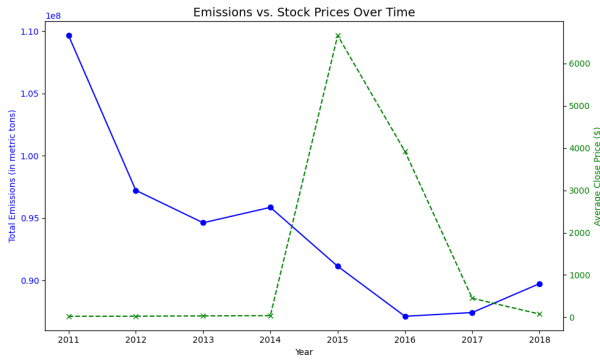


Figure 2: Emissions vs Stock Price 2011 to 2018

Data Transformation

To facilitate year-over-year trend analysis, emissions data columns for each year were transformed into a single column format, with a separate "Year" column. This melting process allowed longitudinal analysis and easier cross-referencing with annual financial metrics.

The final merged Finance and Emission Data enables multi-year, cross-sectional analyses of emissions data alongside financial performance indicators. By combining NASDAQ stock data with GHGP facility emissions, this study supports investigations into potential correlations between environmental emissions and corporate financial performance.

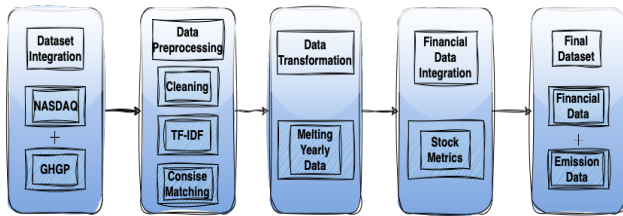


Figure 3: Finance and Emission Data Creation

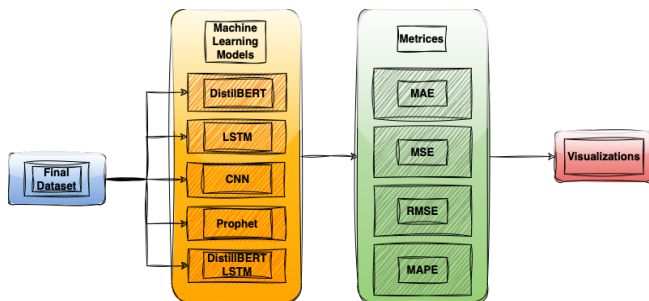


Figure 4: Model Implementation, Metrics and Visualization

Methodology

Model Selection

To predict corporate carbon emissions from financial indicators, we evaluate five distinct models: DistilBERT, LSTM, CNN-LSTM, Prophet and DistilBERT-LSTM. Each model processes four input features: average high price, average low price, average open price, and average close price, aggregated annually per company to predict the corresponding carbon emissions.

DistilBERT: We explore DistilBERT's applicability to time series forecasting, despite its primary design for NLP tasks. This lightweight transformer model retains 97% of BERT's performance on language understanding benchmarks while being 40% smaller and 60% faster during inference. While transformer architectures have shown promising results in various domains, their effectiveness in time series prediction, particularly for carbon emission forecasting, remains relatively unexplored. This investigation aims to assess whether the self-attention mechanisms and compressed knowledge representation of DistilBERT can effectively capture temporal patterns in financial data.

LSTM: Long Short-Term Memory networks serve as our baseline deep learning approach, given their established effectiveness in sequence modeling. The architecture's gated structure, comprising input, forget, and output gates, facilitates the capture of long-term dependencies in temporal data. LSTMs are particularly adept at modeling time series with long-term trends or cycles, making them suitable for our multivariate financial analysis. Their ability to selectively retain or discard information through their memory cells enables effective handling of complex temporal relationships, though this comes with the challenge of careful hyperparameter tuning to achieve optimal performance.

CNN-LSTM: Our CNN-LSTM implementation creates a powerful hybrid architecture that leverages the strengths of both networks. The 1D convolutional layers process local patterns in the financial indicators, automatically extracting hierarchical features from input sequences. These features are then fed into LSTM layers that model temporal dependencies, creating a sophisticated representation of the time series data. This hybrid approach is particularly effective for complex multivariate time series, as it combines CNN's capability to capture spatial features with LSTM's strength in sequence prediction. The architecture excels in handling non-linear relationships and provides more robust predictions compared to single-architecture approaches.

Prophet: We incorporate Facebook's Prophet model to leverage its sophisticated decomposition of time series data. Prophet employs an additive model that integrates non-linear trends with multiple seasonal patterns, including yearly, weekly, and daily seasonality, along with holiday effects. Its robust handling of missing data and outliers makes it particularly suitable for real-world financial data analysis. The model's automatic hyperparameter tuning and capacity to model complex seasonal patterns provide a strong benchmark for our evaluation. While Prophet traditionally excels in long-term predictions with substantial historical data, we

DistilBERT-LSTM: The research uses a novel hybrid machine learning model that combines DistilBERT’s natural language processing skills with LSTM’s time series analysis approaches. By using a pre-trained DistilBERT model for feature embedding, followed by an LSTM layer to capture temporal relationships and a regression output layer, the approach enables company-specific predictions with granular performance measurement.

Model	MAE ↓	MSE ↓	RMSE ↓	MAPE ↓
DistilBERT	2.796	11.737	2.796	24.271
LSTM	10.856	120.275	10.856	94.658
CNN-LSTM	6.433	44.508	6.433	55.013
Prophet	3.229	104.451	3.229	33.929
DistilBERT-LSTM	8.803	80.316	8.803	76.500

The chart displays four price metrics over time. The 'Average High Price' (blue line) and 'Average Low Price' (orange line) show a significant peak in early 2015, reaching approximately 1750 and 1700 respectively. The 'Average Open Price' (green line) and 'Average Close Price' (red line) follow a similar trend but at lower values, peaking around 1650 and 1600 in early 2015. All four metrics show a general downward trend from 2015 to 2017, with a sharp decline in 2017.0, where the Average High Price drops to around 1500 and the Average Low Price drops to around 1000. The chart then shows a slight recovery in 2018.0, with the Average High Price rising to around 1600 and the Average Low Price rising to around 1100.

Year	Average High Price	Average Low Price	Average Open Price	Average Close Price
2014.0	1000	800	900	850
2014.5	1200	1000	1100	1050
2015.0	1750	1700	1650	1600
2015.5	1500	1400	1450	1400
2016.0	1300	1200	1250	1200
2016.5	1100	1000	1050	1000
2017.0	1500	1000	1400	1350
2017.5	1400	1100	1300	1250
2018.0	1600	1100	1500	1450

We implement a company-wise training approach where each model is trained independently on individual company

Figure 6 presents a comparative graph of carbon emission predictions by DistilBERT against ground truth values for

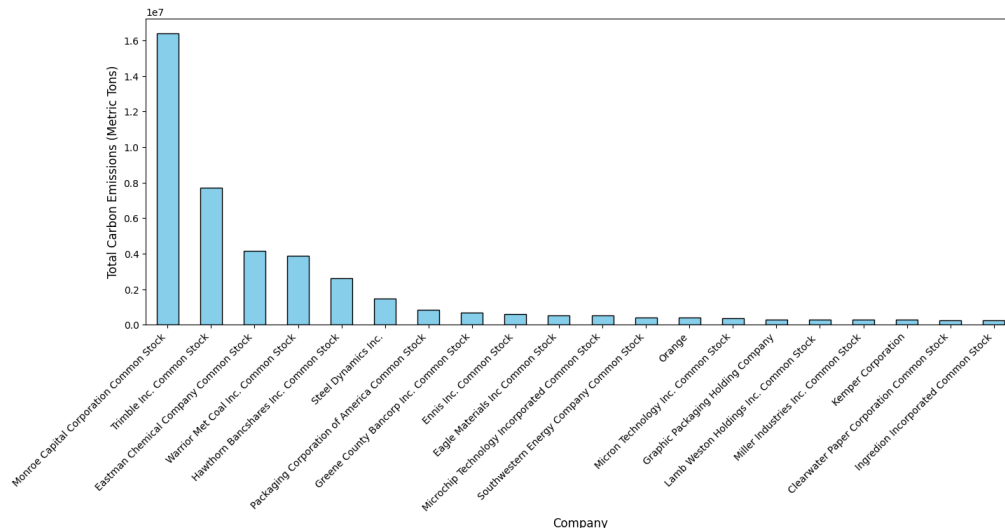


Figure 7: Carbon Emissions of 20 Companies in 2018

50 companies. The graph represents carbon emission values using logarithmic transformation of the original data. Fig 7 represents the top carbon-emitting companies based on the dataset, depicting the large environmental influence of some organizations. Monroe Capital Corporation has the largest carbon emissions, topping 1.6×10 metric tons. This severe gap emphasizes the importance of targeted actions in industries with disproportionately high emissions in order to make meaningful progress towards sustainability targets. Fig 5 displays the average stock values for these firms from 2014 to 2018. Understanding the link between company market performance and its environmental impact is critical to determining whether financially successful organizations pursue sustainable practices or whether profitability correlates with increased emissions due to operational scale. Companies can use the finance and emission data and the predictive analysis to design strategies for decreasing their environmental effect. By accurately forecasting emissions, organizations can optimize resource allocation and improve operational efficiency. These insights motivate businesses to invest in cleaner technologies and sustainable practices. Companies can use data-driven methods to make informed decisions that support ecological responsibility and drive technological innovation in sustainability.

Conclusion

In this paper, we propose a novel dataset containing financial and emission data, including annual information on various companies' stock prices and carbon emissions. We used this data to train different deep learning models that forecast carbon emissions based on stock values. Among the 5 models, DistillBERT gave the best results with MAE of 2.796, MSE of 11.737, RMSE of 2.796 and MAPE of 24.271. This study can help make a paradigm shift towards evidence-based sustainability by incorporating data and using deep learning models for analysis. Companies can move from traditional compliance-driven methods and adopt a more proactive and

anticipatory approach to environmental governance. This approach can help in reducing carbon emissions at global level.

Limitation: We based our study on each company's average stock value throughout the course of the year. While this strategy provides a broad perspective, using daily stock data could improve our models' accuracy by capturing year-round changes and trends. Furthermore, the emission data used in this study is from 2011 to 2018, which does not account for more recent trends and advancements. These limitations can be addressed in future work.

References

- Adamolekun, G. 2024. Carbon price and firm greenhouse gas emissions. *Journal of Environmental Management*, 349: 119496.
- Antoniuk, Y.; and Leirvik, T. 2024. Climate change events and stock market returns. *Journal of Sustainable Finance & Investment*, 14(1): 42–67.
- Aswani, J.; Raghunandan, A.; and Rajgopal, S. 2023. Are Carbon Emissions Associated with Stock Returns?*. *Review of Finance*, 28(1): 75–106.
- Box, G. E.; and Jenkins, G. M. 1976. *Time series analysis: forecasting and control*. Holden-Day.
- Chang, C.-L.; Ilomäki, J.; Laurila, H.; and McAleer, M. 2020. Causality between CO2 Emissions and Stock Markets. *Energies*, 13(11).
- Chen, H.; and Wang, L. 2020. Financial indicators and environmental performance: Evidence from mining companies. *Sustainability*, 12(10): 4067.
- Chollet, F. 2017. *Deep learning with Python*. Manning Publications.
- Eccles, R. G.; Ioannou, I.; and Serafeim, G. 2014. The impact of corporate sustainability on organizational processes and performance. *Management Science*, 60(11): 2835–2857.

- Eccles, R. G.; and Serafeim, G. 2013. The performance frontier: Innovating for a sustainable strategy. *Harvard Business Review*, 91(5): 50–60.
- Friede, G.; Busch, T.; and Bassen, A. 2015. ESG and financial performance: aggregated evidence from more than 2000 empirical studies. *Journal of Sustainable Finance & Investment*, 5(4): 210–233.
- Giovanna Bua, F. R., Daniel Kapp; and Rognone, L. 2024. Transition versus physical climate risk pricing in European financial markets: a text-based approach. *The European Journal of Finance*, 30(17): 2076–2110.
- Goodfellow, I.; Bengio, Y.; and Courville, A. 2016. *Deep learning*. MIT Press.
- Graves, A. 2012. *Supervised sequence labelling with recurrent neural networks*. Springer.
- Hengge, M.; Panizza, U.; and Varghese, R. 2023. Carbon Policy and Stock Returns: Signals from Financial Markets. *IMF Working Papers*, 2023(013): A001.
- Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural Computation*, 9(8): 1735–1780.
- Hyndman, R. J.; and Athanasopoulos, G. 2018. *Forecasting: principles and practice*. OTexts.
- International Energy Agency. 2021. Global Energy Review: CO2 Emissions in 2020. Technical report, IEA.
- IPCC. 2018. Global Warming of 1.5°C. Technical report, Intergovernmental Panel on Climate Change.
- McKinsey Global Institute. 2020. Climate risk and response: Physical hazards and socioeconomic impacts. Technical report, McKinsey & Company.
- Oestreich, A. M.; and Tsiakas, I. 2015. Carbon emissions and stock returns: Evidence from the EU Emissions Trading Scheme. *Journal of Banking Finance*, 58: 294–308.
- Paramati, S. R.; Mo, D.; and Gupta, R. 2017. The effects of stock market growth and renewable energy use on CO2 emissions: Evidence from G20 countries. *Energy Economics*, 66: 360–371.
- Porter, M. E.; and Kramer, M. R. 2011. Creating shared value. *Harvard Business Review*, 89(1/2): 62–77.
- Salehi, M.; Fahimifard, S. H.; Zimon, G.; Bujak, A.; and Sadowski, A. 2022. The Effect of CO2 Gas Emissions on the Market Value, Price and Shares Returns. *Energies*, 15(23).
- Stern, N. 2007. *The Economics of Climate Change: The Stern Review*. Cambridge University Press.
- Tang, J.; and Li, J. 2022. Carbon risk and return prediction: Evidence from the multi-CNN method. *Frontiers in Environmental Science*, 10.
- Tran, V. T.; Phan, D. H. B.; Tee, C.-M.; and Nguyen, D. T. 2024. Unmasking the carbon conundrum: How emissions impact stock price crash risk. *Finance Research Letters*, 64: 105443.
- United Nations. 2020. United Nations Climate Change Annual Report. Technical report, United Nations Framework Convention on Climate Change.
- Wen, F.; Wu, N.; and Gong, X. 2020. China's carbon emissions trading and stock returns. *Energy Economics*, 86: 104627.
- World Economic Forum. 2022. The Global Risks Report 2022. Technical report, World Economic Forum.
- Zafar, M. W.; Zaidi, S. A. H.; Sinha, A.; Gedikli, A.; and Hou, F. 2019. The role of stock market and banking sector development, and renewable energy consumption in carbon emissions: Insights from G-7 and N-11 countries. *Resources Policy*, 62: 427–436.
- Zhang, Y.; Shen, L.; Shuai, C.; Tan, Y.; Ren, Y.; and Wu, Y. 2019. Predictive modeling of carbon emissions based on economic development factors. *Journal of Cleaner Production*, 234: 225–236.