EU Climate Change News Index: Forecasting EU ETS Prices with Online News

Áron Dénes Hartvig[1], Péter Pálos[2], Áron Pap[3]

Highlights

1. The EU Climate Change News Index is created based on the GDELT online news database.[4]
2. The index is used as an alternative predictor for carbon prices.
3. Online news brings predictive power to carbon price forecasting models.

Abstract

Carbon prices have been rapidly increasing in the EU since 2018 and accurate forecasting of EU Emissions Trading System (ETS) prices has become essential. This paper proposes a novel method to generate alternative predictors for ETS prices using relevant online news information. We devise the EU Climate Change News Index by calculating the term frequency-inverse document frequency (TF-IDF) feature for climate change-related keywords. The index is capable of tracking the ongoing debate about climate change in the EU. Finally, we show that incorporating the index in a simple predictive model significantly improves forecasts of ETS prices.

JEL Classification: C80, Q48, Q54, Q58
Keywords: Emissions trading system, Carbon price prediction, Online news, TF-IDF, Climate change, Market index

[1]Corresponding author at: Corvinus University of Budapest, Cambridge Econometrics, Fővám square 8., Budapest, 1093, Hungary
[2]Independent scholar, Budapest, Hungary
[3]Independent scholar, Budapest, Hungary
[4]The Global Database of Events, Language, and Tone (GDELT) Project is a real-time network diagram and database of global human society for open research that monitors the world's broadcast, print, and web news in over 100 languages. For more information, see: https://www.gdeltproject.org/, accessed: 2022-10-21.

1. Introduction

Carbon pricing is an economically efficient instrument in the policy toolkit to signal emissions' actual cost and stimulate investments in low-carbon technological innovations. Emissions trading systems are market-based mechanisms where a cap is set on certain sectors' emissions and the entities covered are allowed to trade emissions allowances. In theory, trading ensures that emissions are reduced where abatement costs are the lowest. The world's first international carbon market, the EU Emissions Trading System (ETS), was implemented in 2005 covering power generators and energy-intensive industries. ETS prices stayed low for a long time but finally started to rise in 2017. The ETS allowance prices skyrocketed in 2021, reaching almost €100 in February 2022. Nevertheless, prices dropped below €60 in early-March after the disruptions caused by the Russian invasion of Ukraine.

The current price range and the peak price of almost €100 have caught the eyes of investors. However, volatility of the allowance prices and policy uncertainty around the future of the ETS increases risks associated with low-carbon technology investments. Furthermore, most companies covered by the ETS are risk averse in terms of trading with allowances [10] and volatile carbon prices increase the uncertainty of their production costs since they purchase allowances as they emit.

Our contribution to the literature on carbon pricing is twofold. First, we propose a new EU Climate Change News Index (ECCNI) that tracks the ongoing discussion in the EU about climate change using global media sources. Although some articles have already incorporated news information into the forecasting of ETS prices [17, 18], no meaningful feature has been created that captures the media's climate policy-related discussions. We apply term frequency–inverse document frequency (TF-IDF) feature extraction to the GDELT[5] news database to measure the frequency of climate change-related keywords in the news associated with the EU. The TF-IDF features help to quantify the intensity of the discussion about climate change in the EU and to incorporate policy context in the analysis of carbon prices.

Second, we apply the news index to predict the next day's ETS allowance price returns. We test the forecast accuracy of the ECCNI against a set of control variables taken from the literature. Our results suggest that the occurrence of climate change-related keywords in the most reliable news sites improves the forecasts of the ETS prices.

Several quantitative methods have already been developed in academic literature to forecast carbon prices. Zhao et al. [19] categorizes these papers into two groups: (1) forecasting based on time-series data using carbon price only; (2) extending with economic and energy data too. The carbon price-only methods mostly include ARIMA models; however, they can only capture linear relationships [20]. Therefore, more advanced frameworks have been applied to carbon prices, like different varieties of generalized autoregressive conditional heteroscedasticity (GARCH) models [2, 3, 5] and vector autoregressive (VAR) models [2].

Nevertheless, carbon price-only methods do not incorporate all available information in the market. Various articles that aim to forecast carbon prices use economic and energy-related variables proxying the demand for CO_2 allowances [2, 8, 19]. Recently, alternative predictors, e.g., news data through natural language processing (NLP), have also been used to forecast ETS prices. Ye and Xue [17] created a carbon tone index reflecting sentiment in news articles and showed that it has strong predictive power on carbon prices. Zhang and Xia [18] used online news data and Google trends to improve the forecast of ETS prices. However, they only included the headlines and titles of online news from limited sources and consequently could only examine carbon prices with a weekly frequency. Furthermore, their word embedding algorithm is not directly interpretable and lacks transparency. To shed more light on the impact of news on daily ETS prices, we create features by applying the TF-IDF method to the GDELT news dataset. TF-IDF has been widely used to improve the forecast accuracy of stock prices [6, 12, 14, 15].

The remainder of this paper is organized as follows. Section 2 provides a short description of our dataset, while Section 3 outlines the methods used to analyse it. This leads into Section 4, where we discuss the generated ECCNI and the forecasting performance of the models that incorporate news information. Finally, Sections 5 and 6 summarise our conclusions and ideas for future work.

[5]The Global Database of Events, Language, and Tone (GDELT) Project is a real-time network diagram and database of global human society for open research that monitors the world's broadcast, print, and web news in over 100 languages. For more information, see: https://www.gdeltproject.org/, accessed: 2022-10-21.

2. Data

The main contribution of this study is the conversion of online news articles to meaningful variables that enhances our understanding of ETS. Therefore, we use GDELT, a free open platform covering global news from numerous countries in over 100 languages with daily frequency. The database includes, along with others, the actors, locations, organizations, themes, and sources of the news items [11]. GDELT has been used in various articles that apply NLP to extract alternative information from the news [1, 7, 9].

We take the daily futures closing prices of the European Union Allowance (EUA) (€/ton) as the dependent variable since that is the underlying carbon price of ETS. Besides news data, we include the most fundamental drivers of ETS prices [17] in our analysis to serve as control variables:

- Gas: NBP Natural Gas Futures (NGLNMc1)
- Electricity: Electricity Yearly Futures (ELCBASYc1)
- Coal: Rotterdam Coal Futures (ATWMc1)
- Oil: Brent Oil Futures (LCOF3)
- Stocks: Europe 600 Index (STOXX).

The data was collected from January 2, 2018 until November 30, 2021, with 1011 daily observations in total. The availability of control variables gave the starting date, and the end was determined to avoid the possible distorting effect of the Russian-Ukrainian conflict.

3. Methodology

This study suggests creating the ECCNI and incorporating this index into the forecasts of ETS prices. The following subsection outlines the methodology of the index construction.

3.1. Article collection

Our ECCNI relies on the GDELT database that gathers a wide range of online news with daily frequency. Thus, to focus our analysis, we restricted the dataset to the articles where the actor is *European Union* or *EU* and extracted their URL-s. We chose to filter on the actor to focus on issues and policies that are dealt with by the EU. Moreover, carbon prices are also affected by global trends; consequently, filtering based on geography would not be adequate.

Moreover, we removed the articles from the database that were coming from unreliable sources. For this purpose, we used one of the most cited media bias resources, Media Bias Fact Check (MBFC) [13]. We removed the articles from the data that appeared on 'questionable' websites according to the 'Factual/Sourcing' category of MBFC[6].

After the filtering, the overall number of news sites was reduced from 9,497 to 719, from which our web scraper collected 27,777 articles.

3.2. Feature generation workflow

We performed basic string pre-processing steps on the raw texts using the Natural Language Toolkit (NLTK) package [4]. This package was also used to lemmatize words with WordNetLemmatizer, a more advanced solution than standard stemming because of the addition of morphological analysis. Since our keyword collection contains several multi-word elements, bigrams and trigrams were also formed with the lemmatizer to create the Term Frequency-Inverse Document Frequency (TF-IDF) matrix, which is one of the most commonly used methods for NLP. The TF-IDF method is an adequate tool to incorporate alternative information to forecast financial time series [6, 12, 14, 15]. It is generally accepted that the normalized form of TF-IDF is more effective than Bag-of-words methods in terms of ignoring common words, and it is also able to highlight rare terms.

[6]We are grateful to Courtney Pitcher who fetched the data from MBFC and published an organized dataset on her blog [16].

The rows of our calculated matrix represent the individual articles, and its columns are the elements of the partially external, partially custom-defined keyword list. We gathered our keywords around five main groups: fossil fuels, renewable energy carriers, energy policy, emissions and gas as an independent topic. We used keyword suggestions from Google Trends and our intuition to expand the mentioned groups. The complete list of keywords is shown in Table A.1. We calculated the score for each keyword so it can also be used for further detailed analysis. Still, due to the high variance of the occurrences and the strong correlation between the keyword groups (shown in Figure A.1), we created the EU Climate Change News Index as the aggregated TF-IDF score of the groups[7].

3.3. Forecasting models

The first (*TF-IDF*) model includes the lags of the ETS price returns[8] (r_t) and the ECCNI (z_t) as predictors:

$$r_t = c + \sum_{i=1}^{k} \Big(\phi_i \, r_{t-i} + \theta_i \, z_{t-i} \Big). \tag{1}$$

While the second model, called *Control*, serves as a benchmark model which considers the lags of the ETS price returns and the fundamental driving factors of carbon prices based on academic literature (lags of gas, electricity, coal, oil and stock price returns represented by vector x_t for period t):

$$r_t = c + \sum_{i=1}^{k} \Big(\phi_i \, r_{t-i} + \beta_i^T \, x_{t-i} \Big). \tag{2}$$

The final, *Full* model includes all predictors: the lags of the ETS price returns, the control variables' price returns and the ECCNI:

$$r_t = c + \sum_{i=1}^{k} \Big(\phi_i \, r_{t-i} + \beta_i^T \, x_{t-i} + \theta_i \, z_{t-i} \Big). \tag{3}$$

We use OLS regression and ElasticNet shrinkage method[9] to estimate the models.

4. Results

In this section, we present the ECCNI, the index constructed from the TF-IDF features that are derived using our methodology described in Section 3. First, we assess the evolution of the index qualitatively by walking through the most important events related to climate change in the EU since 2020. We then use OLS and ElasticNet models to test the forecasting ability of the index.

4.1. EU Climate Change News Index

Since policy uncertainty is substantial around the ETS system, measuring the intensity of debate around it is crucial. One of the key drivers of the ETS prices is the EU's ever-increasing emissions reduction targets, which set a cap on the number of ETS allowances. Nevertheless, various other policy measures also impact carbon prices as sectoral policies, like green energy mandates.

We present the evaluation of the ECCNI between January 2, 2018, and November 30, 2021, in Figure 1. The index is highly volatile, but several cycles are outlined in the 30-days moving average. In the followings, we concentrate on the events influencing the index starting from 2020. In January and February 2020, the index reached a relatively high and stable level due to the recent presentation of the EU Green Deal. Then, in March 2020, the index started to decrease steadily as the COVID-19 pandemic overtook the public discourse

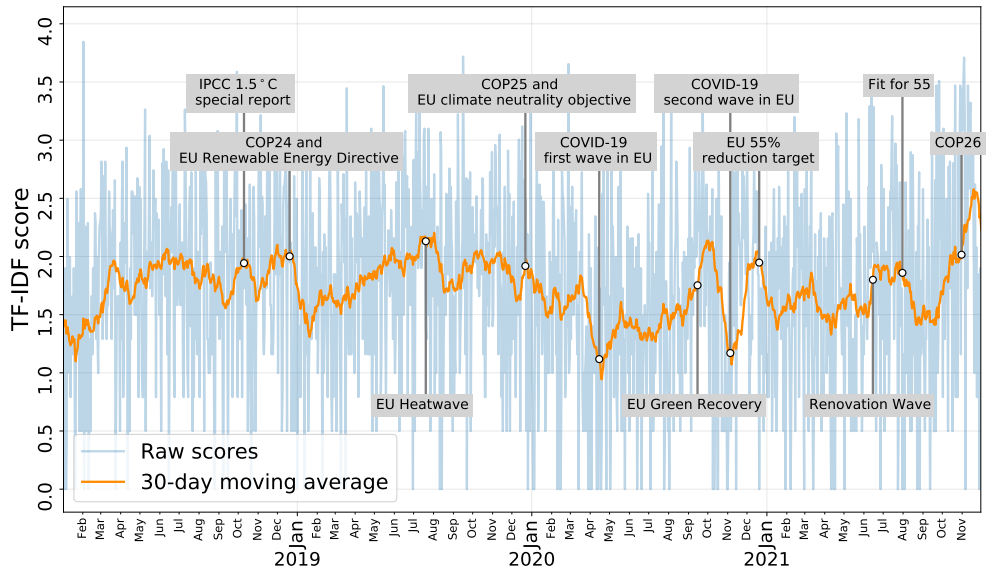
[7] The TF-IDF scores of the EU Climate Change News Index and the keyword groups is available on the EU ETS news tracker dashboard

[8] By price return of variable p we mean the log return: $\Delta log(p_t) = log(\frac{p_t}{p_{t-1}})$

[9] We used the following hyperparameters for grid search: $L1$ = [0, 0.01, 0.05, 0.1, 0.3, 0.5, 0.7, 0.9, 0.95, 0.99, 1] and α = [0, 0.001, 0.003, 0.005, 0.007, 0.009, 0.0095, 0.01, 0.1, 0.3, 0.5, 0.7, 0.9, 0.95, 0.99, 0.999, 1]

in the EU. However, the concept of 'green recovery' soon emerged, and climate change keywords again started trending. In October 2020, COVID-19 cases soared again, pushing down the index. The European Council endorsed a binding EU target of a net domestic reduction of greenhouse gas emissions by at least 55% by 2030 (compared to 1990 levels), leading to a local peak at the end of the year. The proceeding period was less volatile; however, in June 2021, the index jumped to a higher level as the European Council endorsed the new Renovation Wave strategy, and in July, the 'Fit for 55' package was presented. Finally, the climate change news index peaked in November 2021 as the 26th Conference of the Parties (COP26) was held between October 31 and November 13 in Scotland.

Figure 1: EU Climate Change News Index between January 2, 2018 and November 30, 2021



4.2. Forecasting performance

We argue that the EU Climate Change News Index can track the ongoing discussion about the ETS. Since ETS prices are strongly dependent on the policy environment and the measures introduced, the index could potentially help to better predict the evolution of carbon prices in the EU. Therefore, in the followings, we test the forecasting performance of the index. In our analysis, we compare three models to measure the forecasting performance of the ECCNI. Only the lagged values of the predictors are included in the models to produce forecasts that rely entirely on historical information. We run the models with $k = 1, 2, 3, 4, 5$ lags for robustness purposes but only report the results from the best-performing models. Table 1 summarises the out-of-sample 1-day ahead forecast results (MAE and $RMSE$) of the *TF-IDF*, *Control* and *Full* models for carbon price return with different test windows. We used the last $n \in \{50, 75, 100\}$ days of the sample for the out-of-sample testing to examine the performance of the models on the most recent data.

Based on the results, the *Full* model consistently outperforms the others regardless of the test window, the evaluation metric and the estimation method, while the *TF-IDF* model produces the largest errors. These outcomes are in line with the literature exploring the effectiveness of additional textual information in carbon price prediction [17, 18]. News information alone cannot outperform the control variables, but extending these fundamental driving factors with the ECCNI provides additional predictive power to ETS price forecasting. The ECCNI captures policy uncertainty and is able to track the discussion about climate change in the EU.

Table 1: Forecast performance comparison of different models (10^{-3})

Test window	Measure	Model	TF-IDF	Control	Full
50	MAE	OLS	20.22891	18.98009	**18.91033**
		ElasticNet	20.50418	20.25187	**20.14700**
	RMSE	OLS	0.76571	0.69012	**0.68978**
		ElasticNet	0.78675	0.74385	**0.74026**
75	MAE	OLS	18.12085	17.27791	**17.18600**
		ElasticNet	18.27116	18.00239	**17.88948**
	RMSE	OLS	0.64362	0.60436	**0.60363**
		ElasticNet	0.65887	0.62647	**0.62402**
100	MAE	OLS	17.39619	16.67049	**16.57782**
		ElasticNet	17.48796	17.34786	**17.29831**
	RMSE	OLS	0.57801	0.54112	**0.54108**
		ElasticNet	0.58773	0.55749	**0.55667**

5. Conclusions

In this paper we first aggregated textual information from online news articles representing a novel data source for carbon price prediction. We produced TF-IDF features tracking the relative occurrences of climate change-related keywords in online news related to the EU. Then, we derived the EU Climate Change News Index as the aggregated TF-IDF score of the keywords. The index accurately reflects the ongoing discussion about climate change in the EU. It outlines the most influential events in the topic like the annual United Nations Climate Change Conferences or the endorsement of the EU's emissions reduction target of at least 55% by 2030 below 1990 levels. Finally, we showed that the index brings valuable additional information and predictive power to ETS price forecasting compared to a control model where the traditional predictors of carbon prices are included.

The increasing ambition of the EU climate targets brings significant uncertainty to carbon prices. ETS market participants are ever more exposed to the rapidly changing carbon prices; hence, news articles about EU climate issues are highly relevant to their market expectations. Therefore, the proposed ECCNI could also help to manage volatility in the EU ETS. By integrating the index into forecasting models, companies can predict ETS prices more accurately and lower their associated risks.

6. Further research

The NLP method applied in our study offers a great variety of further research ideas. First, ECCNI could be incorporated into more sophisticated forecasting models. Second, in this study, we presented the aggregated time series of the climate keywords. Other groupings or individual TF-IDF time series can be further used in other analyses, like forecasting natural gas prices. Also, our current models assume that the impact of the different variables always has the same direction; however, this assumption could be relaxed and explored with Markov-switching models. Finally, the TF-IDF matrices, including the frequency of the keywords by article, could provide the basis for any news network analysis where the nodes are keywords, and the edges are the number of articles in which both keywords appear. This analysis could help to better understand the interconnections between climate change topics in the media. Furthermore, other NLP methods could be explored with larger datasets, for example *word embeddings* models to improve the representation of contextual information from the news articles.

Statement of exclusive submission

This paper has not been submitted elsewhere in identical or similar form, nor will it be during the first four months after its submission to the Publisher.

Declaration of Competing Interest

Declarations of interest: none

Acknowledgments and funding information

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] Rawan Alamro, Andrew McCarren, and Amal Al-Rasheed. Predicting Saudi stock market index by incorporating GDELT using multivariate time series modelling. *International conference on computing*, pages 317–328, 2019. doi: 10.1007/978-3-030-36365-9_26.

[2] Mohamed El Hédi Arouri, Fredj Jawadi, and Duc Khuong Nguyen. Nonlinearities in carbon spot-futures price relationships during Phase II of the EU ETS. *Economic Modelling*, 29(3):884–892, 2012. doi: 10.1016/j.econmod.2011.11.003.

[3] Thijs Benschopa and Brenda López Cabreraa. Volatility modelling of CO2 emission allowance spot prices with regime-switching GARCH models. SFB 649 Discussion Paper No. 2014-050, Humboldt University of Berlin, Collaborative Research Center 649 - Economic Risk, Berlin, 2014.

[4] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc., 2009. doi: 10.1007/s10579-010-9124-x.

[5] Suk Joon Byun and Hangjun Cho. Forecasting carbon futures volatility using GARCH models with energy volatilities. *Energy Economics*, 40:207–221, 2013. doi: 10.1016/j.eneco.2013.06.017.

[6] Scott Coyne, Praveen Madiraju, and Joseph Coelho. Forecasting stock prices using social media analysis. *2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech)*, pages 1031–1038, 2017. doi: 10.1109/DASC-PICom-DataCom-CyberSciTec.2017.169.

[7] Divyanshi Galla and James Burke. Predicting social unrest using GDELT. *International conference on machine learning and data mining in pattern recognition*, pages 103–116, 2018. doi: 10.1007/978-3-319-96133-0_8.

[8] Heia Njóla Gubrandsdóttir and Haraldur Óskar Haraldsson. Predicting the price of EU ETS carbon credits. *Systems Engineering Procedia*, 1:481–489, 2011. doi: 10.1016/j.sepro.2011.08.070.

[9] Massimo Guidolin and Manuela Pedio. Media attention vs. sentiment as drivers of conditional volatility predictions: An application to brexit. *Finance Research Letters*, 42:101943, 2021. doi: 10.1016/j.frl.2021.101943.

[10] Corina Haita-Falah. Uncertainty and speculators in an auction for emissions permits. *Journal of Regulatory Economics*, 49(3):315–343, 2016. doi: 10.1007/s11149-016-9299-1.

[11] Kalev Leetaru and Philip A Schrodt. GDELT: Global data on events, location, and tone, 1979–2012. *ISA annual convention*, 2(4):1–49, 2013.

[12] Arif Ridho Lubis, Mahyuddin KM Nasution, O Salim Sitompul, and E Muisa Zamzami. The effect of the TF-IDF algorithm in times series in forecasting word on social media. *Indonesian Journal of Electrical Engineering and Computer Science*, 22(2):976, 2021. doi: 10.11591/ijeecs.v22.i2.pp976-984.

[13] MBFC. Media bias fact check, 2022. https://mediabiasfactcheck.com/, Accessed: 2022-10-01.

[14] Marc-andre Mittermayer. Forecasting intraday stock price trends with text mining techniques. *Proceedings of the 37th Annual Hawaii International Conference on System Sciences, 2004.*, pages 64–73, 2004. doi: 10.1109/HICSS.2004.1265201.

[15] Azadeh Nikfarjam, Ehsan Emadzadeh, and Saravanan Muthaiyah. Text mining approaches for stock market prediction. *2010 The 2nd international conference on computer and automation engineering (ICCAE)*, 4:256–260, 2010. doi: 10.1109/ICCAE.2010.5451705.

[16] Courtney Pitcher. My pitcher overfloweth, 2019. https://igniparoustempest.github.io/mediabiasfactcheck-bias/, Accessed: 2022-10-12.

[17] Jing Ye and Minggao Xue. Influences of sentiment from news articles on EU carbon prices. *Energy Economics*, 101:105393, 2021. doi: 10.1016/j.eneco.2021.105393.

[18] Fang Zhang and Yan Xia. Carbon price prediction models based on online news information analytics. *Finance Research Letters*, 46:102809, 2022. doi: 10.1016/j.frl.2022.102809.

[19] Xin Zhao, Meng Han, Lili Ding, and Wanglin Kang. Usefulness of economic and energy data at different frequencies for carbon price forecasting in the EU ETS. *Applied Energy*, 216:132–141, 2018. doi: 10.1016/j.apenergy.2018.02.003.

[20] Bangzhu Zhu and Julien Chevallier. Carbon price forecasting with a hybrid ARIMA and least squares support vector machines methodology. *Pricing and forecasting carbon markets*, pages 87–107, 2017. doi: 10.1007/978-3-319-57618-3_6.

7. Appendix

Appendix A. TF-IDF Methdology

TF-IDF measures the importance of a term to a document in a corpus with the formula below:

$$log\,(1 + f_{t,d}) \cdot log\left(1 + \frac{N}{n_t}\right) \tag{A.1}$$

, where

$f_{t,d}$ = raw count of term t in document d,
N = total number of documents in the corpus,
n_t = number of documents containing term t.

For our aggregated and group-level TF-IDF scores we count the appearances of any of the elements in the keyword list or keyword groups (and not just a single term).

Table A.1: TF-IDF keyword list

Group	Keywords
Emissions	carbon dioxide, CO_2, green deal, greenhouse gas, ghg
Fossil fuels	coal, oil, crude, gasoline, diesel, petrol, fuel
Gas	gas
Policy	climate, sustainability, sustainable, environment, ets
Renewables	renewable, electricity, solar power, solar panel, solar energy, wind power, wind turbine, wind energy, nuclear power, nuclear plant, nuclear energy, clean energy, green energy

Figure A.1: Correlation matrix of TF-IDF group scores

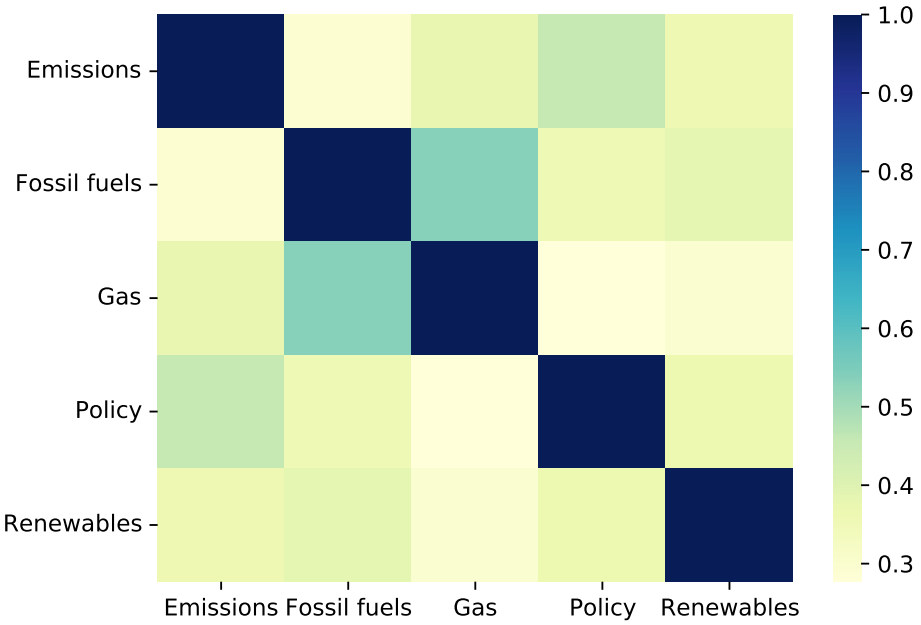
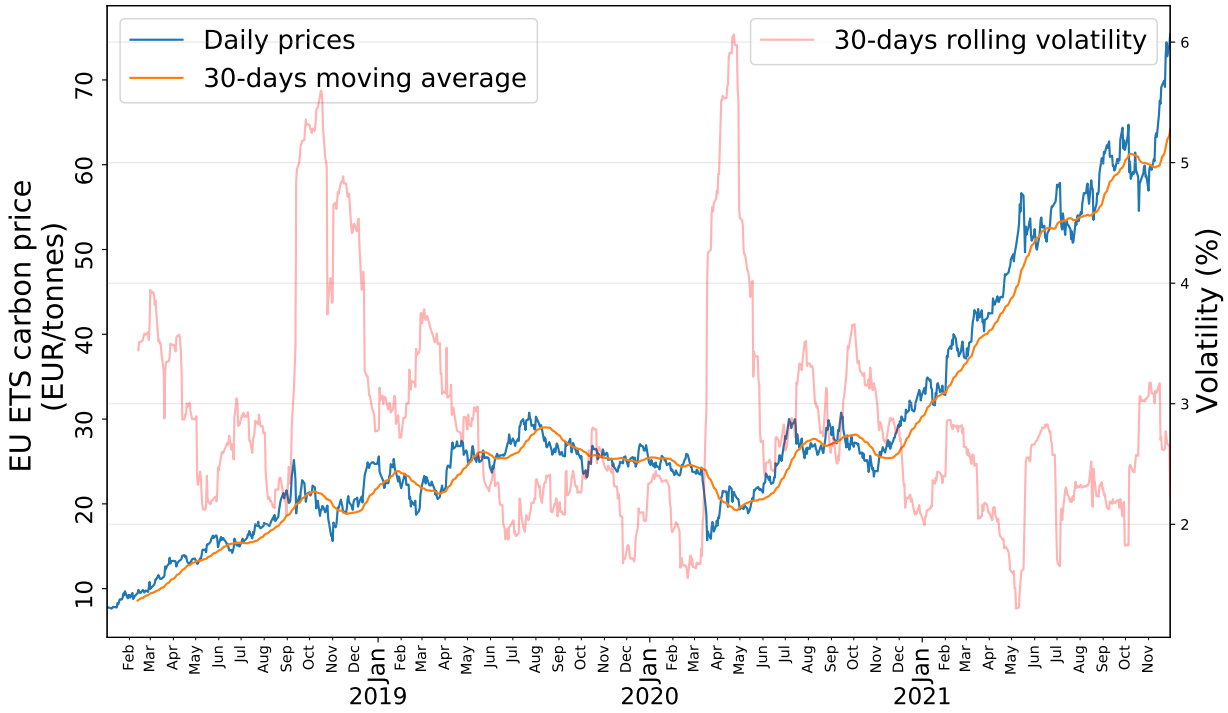
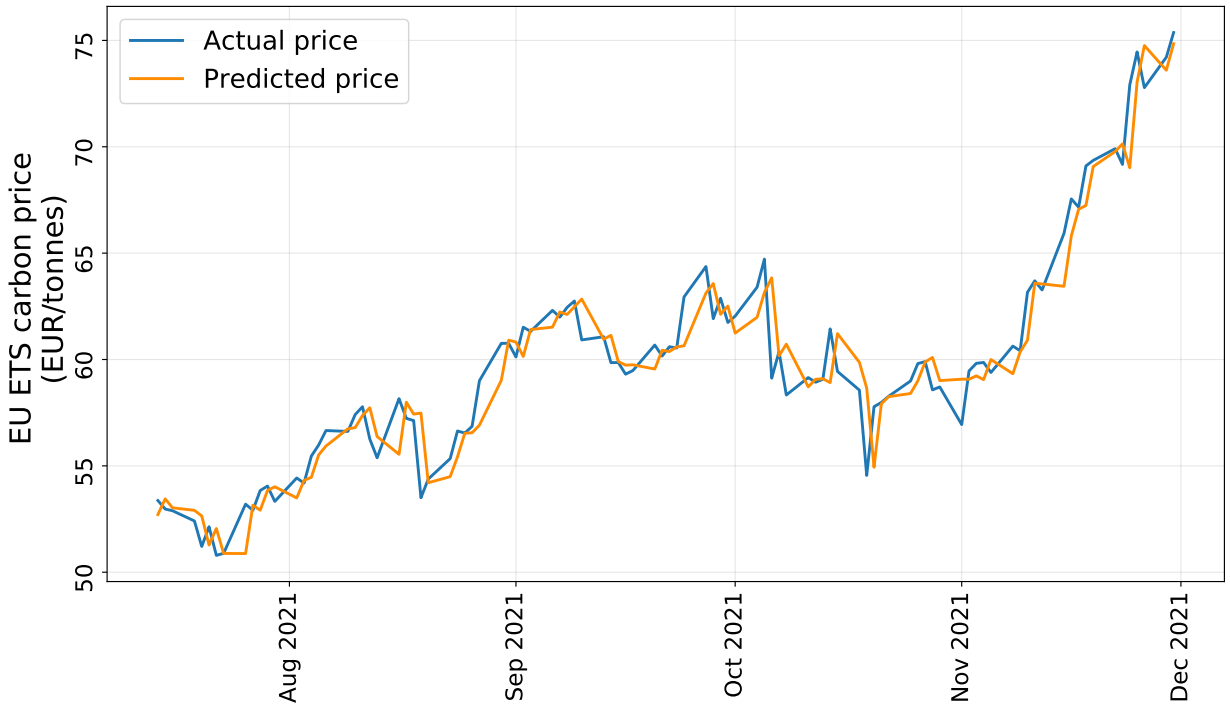


Figure A.2: EU ETS carbon price dynamics between January 2, 2018 and November 30, 2021



Price on the primary axis, volatility on the secondary axis

Figure A.3: Actual and predicted ETS prices in the out-of-sample period from July 14, 2021 to November 30, 2021



Predictions from the full OLS model with 1 allowed lag for the 100-days out-of-sample period