EU climate change news index: Forecasting EU ETS prices with online news

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

Highlights

1. EU climate change news index is composed based on GDELT online news database.
2. The index is used as an alternative predictor for carbon prices.
3. Online news brings predictive power to carbon price forecasting models.

Abstract

The emissions trading system is a key driver of emissions reduction in the EU. Carbon prices have been rapidly increasing since 2020 and accurate forecasting of EU Emissions Trading System (ETS) prices has become essential. In this paper, we propose a novel method to generate alternative predictors for ETS prices using GDELT online news database. We compose the EU climate change news index (ECCNI) by calculating term frequency–inverse document frequency (TF-IDF) feature for climate change related keywords. As climate policies are widely discussed in the news, the index is capable of tracking the ongoing debate about climate change in the EU. Finally, we show that incorporating the ECCNI in a simple predictive model robustly improves forecasts of ETS prices compared to a control model where the traditional predictors of carbon prices are included.

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

[1] Cambridge Econometrics, Corvinus University of Budapest, Fővám square 8., Budapest, 1093, Hungary
[2] Independent scholar, Budapest, Hungary
[3] Independent scholar, Budapest, Hungary

1 Introduction

Carbon pricing is an economically efficient instrument in the policy toolkit to signal the true cost of emissions and to stimulate investments in low-carbon technological innovations. Emissions trading systems are market-based mechanisms where a cap is set on the emissions of certain sectors and the entities that are covered are allowed to trade emissions permits. In theory, trading ensures that emissions reduction takes place where abatement costs are the lowest. The 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 exploded in 2021 reaching almost €100 in February 2022. Nevertheless, prices dropped below €60 in early-March due to the disruptions caused by the Russian invasion in 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 (Haita-Falah, 2016) and volatile carbon prices increase uncertainty of their production costs since they purchase allowances as they emit.

Our contribution to the literature of 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 (Ye and Xue, 2021; Zhang and Xia, 2022), no meaningful feature has been created that captures the climate policy related discussions in the media. We apply term frequency–inverse document frequency (TF-IDF) feature extraction to the GDELT 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 variation in 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 the literature to forecast carbon prices. Zhao et al. (2018) categorizes these papers into two groups: forecasting based on time-series data using carbon price-only, and extending with economic and energy data too. The carbon price-only methods mostly include ARIMA models; however, they can only capture linear relationships (Zhu and Chevallier, 2017). Therefore, more advanced frameworks, like different varieties of generalized autoregressive conditional heteroscedasticity (GARCH) models (Arouri et al., 2012; Benschopa and López Cabreraa, 2014; Byun and Cho, 2013), and vector autoregressive (VAR) model (Arouri et al., 2012) have been applied to carbon prices.

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 the CO2 allowances (Arouri et al., 2012; Gubrandsdóttir and Haraldsson, 2011; Zhao et al., 2018). Recently, alternative predictors, e.g., news data through natural language processing (NLP), have been also used to forecast ETS prices. Ye and Xue (2021) created a carbon tone index reflecting sentiment in news articles and showed that it has a strong predictive power on carbon prices. Zhang and Xia (2022) 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 were only able to examine carbon prices with 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 TF-IDF method to GDELT news dataset. TF-IDF has been widely used to improve forecast accuracy of stock prices (Coyne et al., 2017; Lubis et al., 2021; Mittermayer, 2004; Nikfarjam et al., 2010).

The remainder of this paper is organized as follows. Section 2 provides a short description of our dataset and 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 incorporating news information. Finally, Section 5 and 6 summarise our conclusions and ideas for future work.

2 Data

One 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 (Leetaru and Schrodt, 2013). GDELT has been used in various articles that apply NLP to extract alternative information from news (Alamro et al., 2019; Galla and Burke, 2018; Guidolin and Pedio, 2021).

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 (Ye and Xue, 2021) in our analysis to serve as control variables:

- Gas
- Electricity
- Coal
- Stocks
- Oil

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

3 Methodology

The aim of this study is to compose the ECCNI and to incorporate 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, the carbon prices are affected by global trends as well; consequently, filtering based on geography would not be adequate.

Moreover, we removed the articles from the database that are coming from unreliable sources. For this purpose we used one of the most cited media bias resource, Media Bias Fact Check (MBFC) (MBFC, 2022). We removed the articles from the data that appeared on 'questionable' websites according to the 'Factual/Sourcing' category of MBFC[4].

After the filtering the overall number of news sites 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 (Bird, 2009). This package was also used to lemmatize words with WordNetLemmatizer, which is 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 TF-IDF matrix. The rows of our calculated matrix represent the individual articles, and its columns are the elements of the keyword list.

Term Frequency — Inverse Document Frequency (TF-IDF) is one of the most commonly used methods for NLP.

The per article TF-IDF scores were calculated not on all terms in the corpus but on the basis of a partially external, partially custom defined keyword list. We gathered our keywords around 5 main groups: fossil and renewable energy carriers, energy policy, emissions and gas as an independent topic. We used keyword suggestions from Google Trends and our own 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, but due to the high variance of the occurrences and the strong correlation between the keyword groups (Figure A.1 in the Appendix), we created a collective TF-IDF matrix for the study.

[4] We are grateful to Courtney Pitcher who fetched the data from MBFC and published an organized dataset on her blog (Pitcher, 2019).

3.3 Forecasting models

The first (*TF-IDF*) model includes the lags of the ETS price returns (r_t) and the ECCNI ($EU\ CCNI_t$) as predictors:

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

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

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

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

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

We use OLS regression and the ElasticNet shrinkage method [5] to estimate the models.

4 Results

In this section we present the TF-IDF features constructed using our methodology described in Section 3. First, we compose an aggregated TF-IDF feature that combines the frequency of all climate change keywords considered in this study. We interpret this TF-IDF feature as an index for the presence of climate change as a topic in EU discussions. We then use OLS and ElasticNet models to test the forecasting ability of the index.

4.1 EU climate change news index (ECCNI)

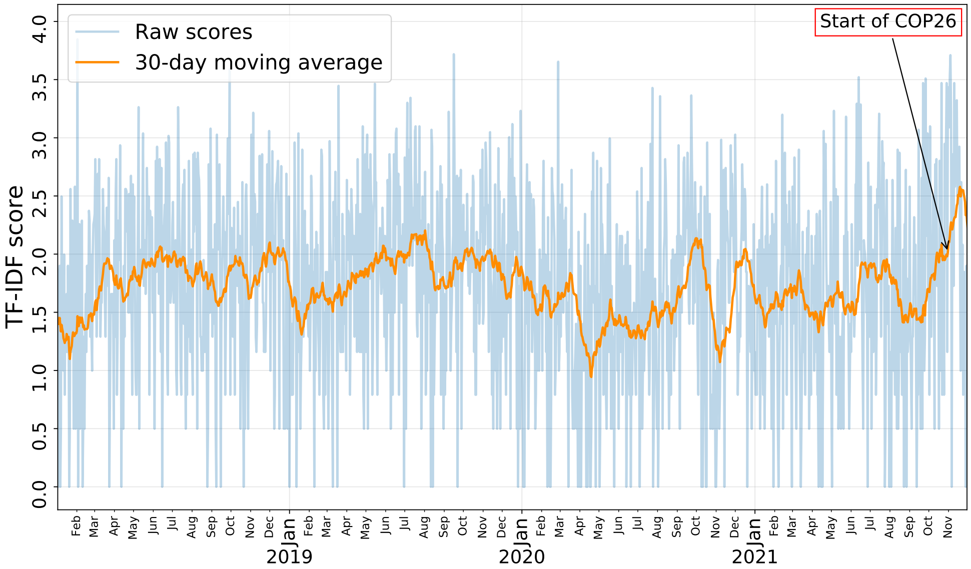
Since policy uncertainty is substantial around the ETS system it is essential to measure the intensity of the discussion around it. One of the key drivers of the ETS prices are the EU's ever increasing emissions reduction targets, which set a cap on the number of ETS allowances. Nevertheless, various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

We present the evaluation of our ECCNI between January 2, 2018, and November 30, 2021, on Figure 1. The index is highly volatile, but several cycles are outlined in the 30-day 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

[5] 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 $\alpha = [0, 0.001, 0.003, 0.005, 0.007, 0.009, 0.0095, 0.01, 0.1, 0.3, 0.5, 0.7, 0.9, 0.95, 0.99, 0.999, 1]$

stable level due to the recent presentation of the EU Green Deal. Then, in March 2020 the index started to steadily decrease as the COVID-19 pandemic overtook the discussion in the EU. However, soon the concept of 'green recovery' emerged and climate change keywords again became trending. In October 2020, the 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 in December leading to a local peak at the end of the year. The proceeding period was less volatile, the index remained in the range of 1.32 and 1.85. Then, the index broke out of the range in June 2021 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 climate change news index is capable of tracking 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 the carbon prices in the EU. Therefore, in the followings we test the forecasting performance of the index. The TF-IDF method is an adequate tool to incorporate alternative information to the forecast of financial time series (Coyne et al., 2017; Lubis et al., 2021; Mittermayer, 2004; Nikfarjam et al., 2010). 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

Table 1: Forecast results (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**

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. This result is in line with our expectations as most of the control variables directly affect the emissions of the companies covered by the ETS while the news index captures policy uncertainty. For example, energy price changes can alter the merit order of power plants. Nevertheless, the ECCNI holds supplementary information to the traditional factors and extending the forecasting model with the news index helps to reduce uncertainty around the ETS prices.

5 Conclusions

In this paper we first aggregated textual information from online news articles which represents a novel data source for carbon price prediction. Then, we derived the ECCNI using TF-IDF methodology that is able to track the ongoing discussion about climate change in the EU. Finally, we showed that the index brings valuable additional information and predictive power to ETS price forecasting models. The ETS market participants are ever more exposed to the rapidly increasing carbon prices; hence, news articles about EU climate issues are highly relevant for their market expectations. Therefore, the proposed ECCNI could also help to manage EU ETS volatility. By integrating the index into forecasting models the 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 analysis, like forecasting natural gas prices. Also, our current models assume that the impact of the different variables always have 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 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 the topics of climate change in the media.

References

Alamro, R., McCarren, A., and Al-Rasheed, A. (2019). Predicting saudi stock market index by incorporating gdelt using multivariate time series modelling. In *International conference on computing*, pages 317–328. Springer.

Arouri, M. E. H., Jawadi, F., and Nguyen, D. K. (2012). Nonlinearities in carbon spot-futures price relationships during phase ii of the eu ets. *Economic Modelling*, 29(3):884–892.

Benschopa, T. and López Cabreraa, B. (2014). Volatility modelling of co2 emission allowance spot prices with regime-switching garch models. Technical report, SFB 649 Discussion Paper.

Bird, Steven, E. L. E. K. (2009). Natural language processing with python.

Byun, S. J. and Cho, H. (2013). Forecasting carbon futures volatility using garch models with energy volatilities. *Energy Economics*, 40:207–221.

Coyne, S., Madiraju, P., and Coelho, J. (2017). Forecasting stock prices using social media analysis. In *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. IEEE.

Galla, D. and Burke, J. (2018). Predicting social unrest using gdelt. In *International conference on machine learning and data mining in pattern recognition*, pages 103–116. Springer.

Gubrandsdóttir, H. N. and Haraldsson, H. Ó. (2011). Predicting the price of eu ets carbon credits. *Systems Engineering Procedia*, 1:481–489.

Guidolin, M. and Pedio, M. (2021). Media attention vs. sentiment as drivers of conditional volatility predictions: An application to brexit. *Finance Research Letters*, 42:101943.

Haita-Falah, C. (2016). Uncertainty and speculators in an auction for emissions permits. *Journal of Regulatory Economics*, 49(3):315–343.

Leetaru, K. and Schrodt, P. A. (2013). Gdelt: Global data on events, location, and tone, 1979–2012. In *ISA annual convention*, volume 2, pages 1–49. Citeseer.

Lubis, A. R., Nasution, M. K., Sitompul, O. S., and Zamzami, E. M. (2021). The effect of the tf-idf algorithm in times series in forecasting word on social media. *Indones. J. Electr. Eng. Comput. Sci.*, 22(2):976.

MBFC (2022). Media bias fact check. Accessed: 2022-02-01.

Mittermayer, M.-A. (2004). Forecasting intraday stock price trends with text mining techniques. In *37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the*, pages 10–pp. IEEE.

Nikfarjam, A., Emadzadeh, E., and Muthaiyah, S. (2010). Text mining approaches for stock market prediction. In *2010 The 2nd international conference on computer and automation engineering (ICCAE)*, volume 4, pages 256–260. IEEE.

Pitcher, C. (2019). My pitcher overfloweth. Accessed: 2022-01-12.

Ye, J. and Xue, M. (2021). Influences of sentiment from news articles on eu carbon prices. *Energy Economics*, 101:105393.

Zhang, F. and Xia, Y. (2022). Carbon price prediction models based on online news information analytics. *Finance Research Letters*, 46:102809.

Zhao, X., Han, M., Ding, L., and Kang, W. (2018). Usefulness of economic and energy data at different frequencies for carbon price forecasting in the eu ets. *Applied Energy*, 216:132–141.

Zhu, B. and Chevallier, J. (2017). Carbon price forecasting with a hybrid arima and least squares support vector machines methodology. In *Pricing and forecasting carbon markets*, pages 87–107. Springer.

7 Appendix

A Control variables

- Gas: NBP Natural Gas Futures
- Electricity: Electricity Yearly Futures (ELCBASYc1)
- Coal: Rotterdam Coal Futures (ATWMc1)
- Stocks: Europe 600 Index (STOXX)
- Oil: Brent Oil Futures (LCOU2)

B TF-IDF

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

$$log\left(1+f_{t,d}\right)\cdot log\left(1+\frac{N}{n_t}\right) \tag{4}$$

, where

$$\begin{aligned} f_{t,d} &= \text{raw count of term } t \text{ in document } d, \\ N &= \text{total number of documents in the corpus}, \\ n_t &= \text{number of documents containing term } t. \end{aligned}$$

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. Also, in this modified form, the $f_{t,d}$ and n_t variables apply not only to one term, but to all elements of the keyword list together.

Table A.1: Keyword list

Group	Keywords
Emissions	carbon dioxid, co2, 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

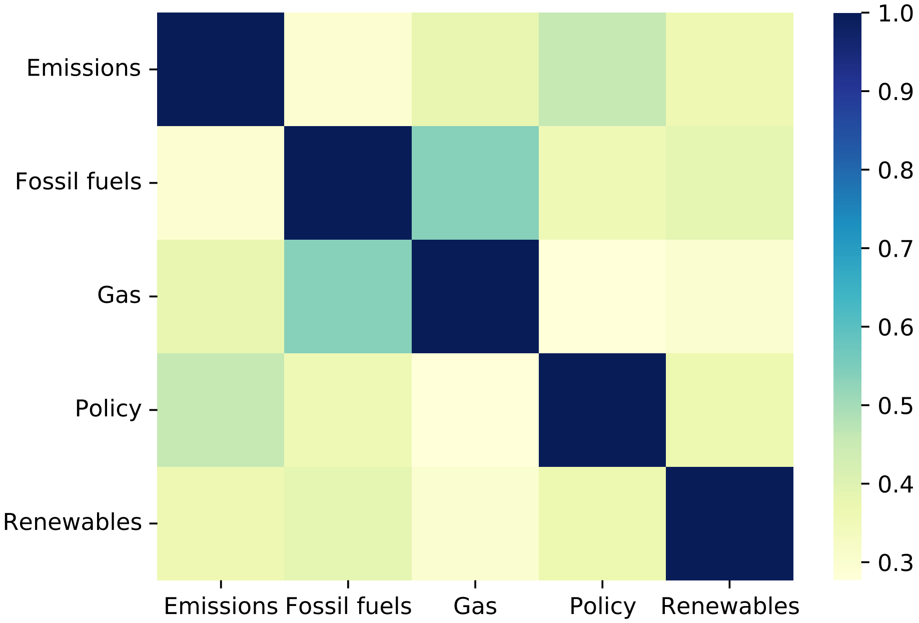
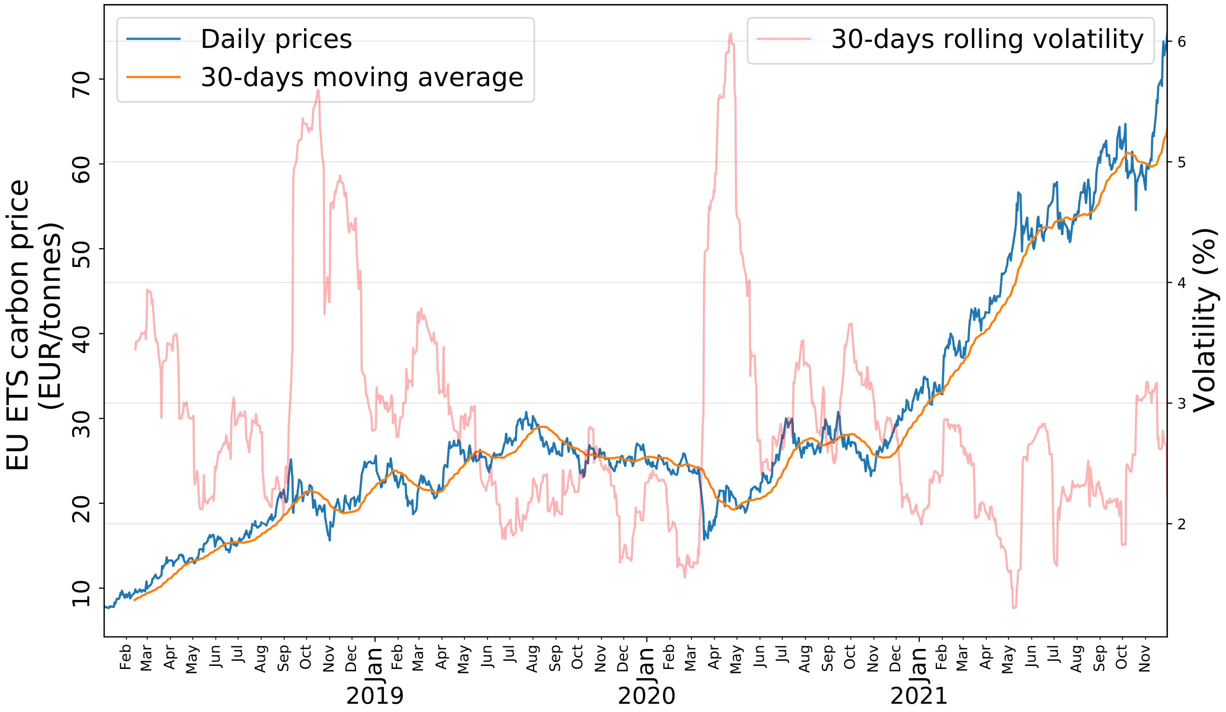


Figure A.2: EU ETS carbon price dynamics in the sample period



Price on the primary axis, volatility on the secondary axis