On the Effectiveness of the EU ETS

BSE Working Paper | 1581 July 2026

Arlet Vila-Bagaria, Jaume Freire-González

bse.eu/research

On the Effectiveness of the EU ETS*

Arlet Vila-Bagaria,† Jaume Freire-González‡

June 30, 2026

Abstract

We evaluate the effectiveness of the European Union Emissions Trading System (EU ETS). Using a staggered difference-in-differences design exploiting variation in adoption timing across countries and phases from 1990–2022, our estimates suggest that the EU ETS is associated with substantial reductions in emissions overall—especially among early participants—although effectiveness varied across cohorts and regions. Later entrants (2008, 2013) showed weaker effects, but disaggregated estimates revealed notable regional reductions. Phase III reforms—tighter caps and expanded auctioning—enhanced performance, cutting emissions by 24MtCO_2 annually by 2020. Overall, the EU ETS has become more effective as its design matured but remains insufficient to meet EU's emissions 2030 targets.

Keywords: EU Emissions Trading System, Carbon markets, Difference-in-differences, Climate policy.

JEL classifications: D62, H23, L50, Q52, Q54, Q58.

* Arlet Vila-Bagaria acknowledges the support of the Utrecht University School of Economics and the PBL Netherlands Environmental Assessment Agency. She further thanks the IDEA program at the Autonomous University of Barcelona for its support. Jaume Freire-González acknowledges financial support from grant PID2024-158997NB-I00, funded by MCIN/AEI/10.13039/501100011033 and ERDF, European Union; from the Severo Ochoa Programme for Centers of Excellence in R&D (Barcelona School of Economics CEX2024-001476-S), funded by MCIN/AEI/10.13039/501100011033; and from AGAUR-Generalitat de Catalunya (2021-SGR-416).

† Utrecht University School of Economics, Utrecht University, 3584 EC, Utrecht, The Netherlands. Email: a.vilabagaria@uu.nl).

‡ Institute for Economic Analysis (IAE-CSIC) and Barcelona School of Economics, Campus UAB, 08193 Bellaterra, Spain. Email: jaume.freire@iae.csic.es.

1. Introduction

Carbon pricing has become a central pillar of global decarbonization strategies, and the European Union Emissions Trading System (EU ETS) remains the world's largest and most influential carbon market. Since its launch in 2005, the EU ETS has evolved through four distinct policy phases, yet its actual effectiveness in reducing greenhouse gas (GHG) emissions remains controversial. This debate reflects substantial heterogeneity in implementation across member states; structural reforms across trading phases; and the complex, staggered adoption timeline that has characterized the program's expansion.

The existing empirical literature has provided valuable insights into the functioning of the EU ETS, but important gaps remain. Most prior studies have focused on individual countries or specific phases of the program, limiting broader conclusions about EU-wide carbon-pricing effectiveness. For example, Petrick and Wagner (2014) examined German manufacturing firms using matching estimators and found a 20% reduction in CO_2 emissions between 2007 and 2010. Jaraite-Kazukauske and Maria (2016) applied a difference-in-differences (DiD) approach to Lithuanian enterprises, also identifying significant emission reductions, whereas Kara et al. (2008) investigated carbon cost pass-through in the Nordic electricity market. Although these studies provided valuable micro-level evidence, their narrow geographical or sectoral scope restricted inferences about the policy's overall impact.

Methodologically, the literature is diverse, and this diversity often produces conflicting results. Synthetic control methods, as in Bayer and Aklin (2020), constructed nuanced counterfactuals by weighting non-ETS sectors. By contrast, studies such as Fowlie et al. (2012) and Fageda and Teixidó-Figueras (2022) relied on conventional DiD frameworks. Other contributions used firm-level matching techniques (Petrick and Wagner, 2014) or macroeconomic modeling approaches (Lise et al. 2010). Yet a notable omission in this literature is the limited use of staggered DiD designs that directly exploit the EU ETS's phased implementation (2005–2007, 2008–2012, 2013–2020, and 2021–2030). This is particularly relevant given the Phase III reforms, including the expansion of auctioning and the introduction of the Market Stability Reserve.

Most empirical work has concentrated on Phase I, with relatively few studies examining later phases using comparable frameworks. Bordignon and Gamannossi degl'Innocenti (2023) provided a key exception, applying DiD to installations across Phases II and III, although their focus remained firm-level rather than sectoral or national. Sector-level evidence is also limited: although some studies have investigated competitiveness concerns (Demailly and Quirion, 2008) or carbon leakage risks (Martin et al. 2014), few have systematically assessed emissions outcomes within the energy sector at the national or EU-wide scale. This gap is notable because energy

production accounts for more than 27% of EU emissions (European Commission, 2025) and forms the core regulatory target of the ETS.

Recent methodological advances have begun to fill some gaps. Biancalani et al. (2024) used machine learning techniques based on matrix completion to estimate an average emissions reduction of 15.4% across EU countries from 2005–2020 although with substantial national variation. However, their study did not differentiate among ETS phases or focus specifically on the energy sector. Similarly, Dechezleprêtre et al. (2023) combined matching and DiD methods to document a 10% reduction in emissions between 2005–2012 without negative competitiveness effects, and Colmer et al. (2025) reported 14%–16% reductions among French manufacturers driven by technological upgrading rather than economic contraction.

Addressing these gaps, we provide a comprehensive assessment of the EU ETS across cohorts and over time, with a focus on the energy sector. Our results suggest a substantial policy effect: Across all specifications, the average treatment effect on the treated (ATT) is roughly ten $MtCO_2$ per year, equivalent to an approximately 19.6% reduction from pre-policy levels. Cohort-level estimates reveal substantial heterogeneity: 2005 entrants showed the largest reductions and the 2008 cohort masked divergent patterns between the Balkans and Nordics, while estimates for the single-country 2013 cohort (Croatia) are reported with caution and are not central to our conclusions. Dynamic effects indicated that reductions began with Phase II, grew steadily, and peaked around 2020 before a slight moderation. These findings remain robust across alternative control groups, covariate adjustments, subgroup analyses, alternative estimators, and a synthetic difference-in-differences design.

We make three main contributions. First, we implemented a staggered DiD framework with heterogeneous treatment effects that explicitly accounts for the phased rollout of the EU ETS across its four policy periods. Second, we conducted a cohort-level analysis of emissions data from 1990–2022 to estimate pan-European effects, moving beyond the firm-level or single-country focus that has dominated prior studies. Third, we focused on the energy sector—the largest source of emissions and the primary target of the ETS—thereby providing clearer evidence of policy effectiveness while avoiding the identification challenges inherent in cross-sectoral analysis.

Our contribution is distinct from and complementary to two recent studies. Dechezleprêtre et al. (2023) exploit installation-level thresholds for ETS participation across four countries (France, Netherlands, Norway, and the United Kingdom) during 2005–2012, focusing on energy-intensive sectors. Colmer et al. (2025) similarly examine Phases I and II using firm-level data to distinguish genuine abatement from outsourcing. Both studies provide credible within-sector identification for the early phases of the ETS. Our paper complements this work along three

dimensions. First, we cover all four regulatory phases through 2022, which is substantively important because the ETS underwent fundamental design changes in Phase III (2013): the shift from national allocation plans to a centralized EU-wide cap, the transition from grandfathering to auctioning as the default allocation method, and significantly tighter allowance limits. These reforms represent a qualitatively different policy regime that cannot be evaluated using data ending in 2012. Second, we include all ETS-participating countries, providing a system-wide assessment of aggregate effectiveness. Third, our cohort-specific estimates reveal important heterogeneity across waves of entrants that is not captured in firm-level studies focused on a subset of countries.

Energy production accounts for roughly 75% of covered emissions, offering a natural experiment of staggered participation across countries and time. We employed a staggered DiD approach following Callaway and Sant'Anna (2021) to address three empirical challenges. First, our approach accounts for heterogeneous treatment effects across cohorts (2005, 2008, 2013), which renders conventional DiD estimators biased, particularly for Phases III and IV. Second, we incorporated covariate adjustment, including renewable energy share, to improve the precision of ATT estimates. Third, we enhanced the robustness of counterfactual comparisons by using both never-treated and not-yet-treated units as controls. This approach harmonized emissions data from multiple sources, constructing a novel dataset covering thirty-seven European countries from 1990 to 2022. To contextualize these results, we first describe the EU ETS, its evolution across four policy phases, and the key structural reforms that shaped emissions incentives. We then outline our empirical strategy to identify the program's impact on energy-sector emissions.

2. The EU ETS

The EU ETS was introduced in 2005 as a cornerstone of the EU's climate policy. It currently involves thirty countries and regulates approximately 36% of the EU's total GHG emissions, equivalent to 1,335 million metric tons of CO_2 equivalent ($MtCO_2e$). The system covers key GHGs, including carbon dioxide (CO_2), hydrofluorocarbons (HFCs), nitrous oxide (N_2O), perfluorocarbons (PFCs), and sulphur hexafluoride (SF_6).

Since its implementation, the EU ETS has contributed to a significant 50% reduction in emissions from electricity and heat generation and industrial manufacturing (European Commission, 2025). It is a critical tool in achieving the EU's ambitious climate goals, which include reducing net GHG emissions to at least 62% below 1990 levels by 2030 and reaching climate neutrality by 2050 (European Commission, 2025). The EU ETS sets a cap on total greenhouse gas emissions by issuing emission allowances, each allowing the release of one tonne of CO_2 equivalent.

These allowances are primarily sold through auctions but can also be traded between companies. As the emissions cap is gradually reduced, the number of available allowances also decreases, creating scarcity.

Revenue generated from the system is primarily allocated to the budgets of member states, with the stipulation that at least 50% of these funds be directed toward climate- and energy-related projects. Additionally, a portion of the revenue can be used to provide financial compensation to industries facing increased costs due to the EU ETS. Companies must report their emissions annually and surrender enough allowances to match them. Failing to do so results in significant fines. Although many allowances are auctioned, some are allocated for free. Companies that reduce their emissions can trade or save unused allowances for future use. The carbon market determines allowance prices and is influenced by strict oversight and the shrinking cap, which helps maintain them. This market-driven price encourages cost-effective emission reductions and has generated revenue, from 2013 onwards, of more than €175 billion. Since its launch in 2005, the EU ETS has evolved through four distinct phases, each with its own characteristics.

2.1. Phase 1: Pilot Phase (2005–2007)

The primary objective of Phase 1 was to test the EU ETS's operational functionality and establish the necessary infrastructure for the subsequent phases. The system initially covered the fifteen EU member states in addition to those that joined in 2004, regulating CO_2 emissions from key sectors including power generation, oil refining, iron and steel production, cement manufacturing, and ceramics (Bordignon and Gamannossi degl'Innocenti 2023). During this phase, 98% of allowances were allocated for free and were primarily based on historical emissions. Firms that exceeded their allocated allowances faced a penalty of €40 per excess tonne of CO_2 (European Commission, 2025). Phase 1 revealed critical design flaws. The total number of allowances issued substantially exceeded actual emissions, creating a market surplus. This overallocation, combined with the inability to bank allowances for future phases, led to a collapse in carbon prices, which fell to nearly zero by 2007 (Ellerman and Buchner 2008). The price volatility and eventual market inefficacy highlighted the need for structural reforms. These lessons led to significant adjustments in subsequent phases, including a gradual shift toward auctioning.

2.2. Phase 2: Kyoto Alignment (2008–2012)

The second phase marked a critical alignment of the EU ETS with the European Union's obligations under the Kyoto Protocol. Several key modifications were implemented to strengthen the system: (1) sectoral expansion to include aviation starting in 2012, (2) incorporation of nitrous

oxide (N_2O) emissions from nitric acid production, (3) a 6.5% reduction in the emissions cap compared with Phase 1, (4) limited permission for the use of international credits from the Clean Development Mechanism and Joint Implementation projects, and (5) a substantial increase in noncompliance penalties from €40 to €100 per excess ton of CO_2. The system's geographic coverage expanded to include Norway, Iceland, Liechtenstein, Bulgaria, and Romania following their EU accession in 2007.

Although these reforms represented significant progress, this phase faced persistent challenges. Reported emissions decreased by approximately 9%, but the problem of overallocation continued to undermine the system's effectiveness. The 2008 financial crisis significantly contributed to emission reductions through decreased industrial output, creating a surplus of allowances and consequent price depreciation in carbon markets. By the conclusion of Phase 2, the EU ETS regulated approximately 38% of the EU's total greenhouse gas emissions, demonstrating its growing centrality in the Union's climate policy.

2.3. Phase 3: Strengthening the System (2013–2020)

This third phase introduced further reforms to address weaknesses identified in earlier phases and to align the system with the EU's long-term climate objectives and expanded to Croatia. The key changes included (1) transitioning from national caps to a single EU-wide cap, declining annually by 1.74%, (2) shifting power-sector allowances from free allocation to auctioning, (3) expanding the coverage to additional industries and new greenhouse gases, and (4) establishing the Market Stability Reserve in 2019. These changes helped to manage surplus allowances and stabilize the market by adjusting the supply of allowances based on predefined rules. They contributed to a greater reduction in emissions and an increase in allowance prices as the caps became more stringent. The EU ETS extended its coverage to sectors such as carbon capture and storage (CCS) and the production of petrochemicals, aluminum, ammonia, and nitric, adipic and glyoxylic acid, further integrating these industries into the emissions trading framework.

2.4. Phase 4: Increased Climate Ambition (2021–2030)

The current phase represents a significant evolution of the EU ETS, aligning with the European Union's enhanced climate ambition to reduce greenhouse gas emissions by 55% below 1990 levels by 2030. Structural reforms characterize this phase, beginning with a more aggressive annual reduction of allowances at 2.2% until 2023, 4.3% from 2024 to 2027, and 4.4% from 2028 onward, nearly doubling the previous phase's linear reduction factor. The Market Stability Reserve has been strengthened through adjusted intervention thresholds to more effectively manage allowance

surpluses and maintain price stability. Notably, the system's scope has expanded to include maritime transport emissions starting in 2024, marking the first major sectoral addition since aviation was incorporated in Phase 2. To address carbon leakage concerns, the innovative Carbon Border Adjustment Mechanism has been introduced, imposing tariffs on imports from jurisdictions with less stringent climate policies. The auctioning share has increased substantially, up to 57%, reflecting a continued transition away from free allocations. These reforms have already had significant market impact, with the average auction price reaching €83.24 per ton of CO_2 in 2023 and generating cumulative revenues exceeding €184 billion since 2013. The current phase also accounts for geopolitical changes, particularly the United Kingdom's withdrawal from the system following Brexit. This enhanced framework positions the EU ETS as a central instrument in achieving neutrality objectives while serving as a model for emerging carbon markets globally. Table 1 shows an overview of the phases.

Table 1—Overview of the EU ETS Phases, Detailing Country Participation, Sector Coverage, Emission Caps and Noncompliance Penalties

	Phase 1	Phase 2	Phase 3	Phase 4
Countries	EU–25	EU–27 + Norway, Iceland, and Liechtenstein	EU–28 + Norway, Iceland, and Liechtenstein	EU–27 (excluding UK) + Norway, Iceland, and Liechtenstein
Sectors	Power generators and energy-intensive industries (oil refineries, coke ovens, iron and steel, cement, glass, lime, bricks, ceramics, pulp, and paper)	Includes nitric acid and aviation	Includes carbon capture and storage, petrochemicals, ammonia, metals, gypsum, aluminum, and nitric, adipic, and glyoxylic acid	Includes maritime sector
Cap	2096 $MtCO_2e$ in 2005 with 90% free allowances	2049 $MtCO_2e$ in 2008 with 90% free allowances	2084 $MtCO_2e$ in 2013, annually reduced by 1.74% until 1816 $MtCO_2e$in 2020; 57% auctioning	From 2021, cap reduced annually: 2.2% (until 2023), 4.3% (2024–2027), 4.4% (from 2028), cuts of 90M (2024), and 27M (2026) allowances, 57% auctioning
Penalty for noncompliance	40€ per tonne	100€ per tonne	100€ per tonne	100€ per tonne

3. Methodology

Isolating the impact of the EU ETS on CO_2 emissions is challenging, not only because of the many variables influencing emissions but also because of the lack of reliable counterfactuals for comparison. Without a clear baseline, attributing observed emissions cuts solely to the EU ETS remains difficult (Laing et al. 2014). To evaluate the policy impact while assessing parallel trends between treatment and control groups, we used data from 1990 to 2022. Although the EU ETS was introduced in 2005 and will continue through at least 2030, our analysis is constrained to the period ending in 2022 due to data availability limitations. The analysis focuses exclusively on CO_2 emissions, which represents the dominant greenhouse gas and accounted for 74.89% of the total global GHG emissions in the world in 2023 and 77.45% in EU–27 for the same year (Jones et al. 2023). This focus is justified by both the policy's primary targeting of CO_2 emissions and the superior data availability for this greenhouse gas compared with other GHGs.

We employed a staggered DiD approach to account for the fact that treatment adoption occurred at different times across countries. A conventional DiD framework would be inappropriate for this analysis because it assumes a single treatment period and homogeneous treatment effects—conditions that clearly do not occur in this setting. Moreover, recent studies have identified significant limitations with both traditional staggered DiD and two-way fixed effects (TWFE) models that are commonly used in this context. Baker and Larcker (2022) showed that TWFE estimators in staggered designs often produce uninterpretable results that may not correspond to ATT, potentially even yielding coefficients with the wrong sign. As Wing et al. (2024) emphasized, TWFE estimators effectively compute weighted combinations of heterogeneous effects, leading to biased estimates when treatment effects vary across time and units (de Chaisemartin and D'Haultfœuille 2020; Goodman-Bacon 2021; Imai and Kim 2021). The robust staggered DiD approach (Callaway and Sant'Anna, 2021) offers three key advantages. First, it provides greater flexibility in control-group selection by incorporating both never-treated units and not-yet-treated units. Second, the method allows for the inclusion of covariates, in our case, renewable energy share, to ensure that the parallel trends assumption holds conditionally and unconditionally. Finally, it allowed us to estimate average treatment effects for each group in each time period while properly accounting for treatment effect heterogeneity. See Appendix III for details on the Callaway and Sant'Anna's (2021) Framework, as well as the identification and empirical strategy applied in this study

Nevertheless, to strengthen the empirical analysis and provide a transparent assessment of our methodological choices, we complement the main Callaway and Sant'Anna (2021) estimates with

a conventional TWFE specification and a Goodman-Bacon (2021) decomposition. The TWFE model serves in this context as a benchmark against which the staggered DiD results can be compared, allowing to evaluate whether accounting for treatment-effect heterogeneity materially alters the estimated impact of the EU ETS. The Goodman-Bacon decomposition further complements the analysis by identifying the specific 2×2 comparisons underlying the TWFE estimate and quantifying the extent to which the coefficient is driven by clean treated-versus-never-treated comparisons or by potentially problematic comparisons involving already-treated units.

3.1. Control and Treatment Groups

Given the staggered design of the EU ETS, identifying the countries within each treatment cohort is straightforward. However, selecting an appropriate never-treated control group is more challenging because it requires identifying countries that are similar to the treated group but have never joined the EU ETS—a difficult task given the policy's near continent-wide coverage. For this analysis, the never-treated control group includes potential future EU entrants from the Western Balkans (Albania, Bosnia and Herzegovina, Montenegro, North Macedonia, and Serbia) and Eastern Europe (Moldova and Georgia). These countries were selected based on their geographic proximity and economic similarities to the treated nations. Ukraine was excluded from this group because of the potential confounding effects of the ongoing military conflict.

During Phase 1 (2005–2007), the treatment group comprised the EU–15 members (Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, Netherlands, Portugal, Spain, Sweden, and United Kingdom) and the ten countries that joined the EU in 2004 (Cyprus, Czech Republic, Estonia, Hungary, Latvia, Lithuania, Malta, Poland, Slovakia, and Slovenia). Additionally, Norway, Iceland, Croatia, Romania, and Bulgaria were included in the not-yet control group in some specifications of the model for this phase because they had not yet implemented the EU ETS.

Phase 2 (2008–2012) maintained the same control group of potential future EU entrants, and the treatment cohort expanded to include Romania and Bulgaria (completing the EU–27) as well as Norway and Iceland, which joined the ETS during this period. Croatia remained in the not-yet control group because it had not yet adopted the policy.

In Phase 3 (2013–2020), the treatment group incorporated Croatia, forming the EU–28, and the control group remained consistent with previous phases. Finally, in Phase 4 (2021–2030), the treatment group corresponding to the 2005 cohort adjusted to the EU–27 following the United Kingdom's withdrawal and the establishment of its independent emissions trading system; the UK

was therefore excluded from this phase.

3.2. Specifications

To ensure robustness and assess the sensitivity of the results, four specifications of Callaway and Sant'Anna's (2021) staggered DiD model were estimated. Model 1 used both never-treated countries (permanent controls) and not-yet-treated countries (future adopters as controls until their treatment date), with no additional covariates. Model 2 maintained the same control group but conditional on renewable share as a covariate. Model 3 used only never-treated controls without controlling for renewable share. Model 4 kept the same control group and added renewable share as covariate.

Given the structural differences between never participating countries and EU ETS member states in terms of differences in economic development, industrial composition, and regulatory environments, we treat specifications relying exclusively on never-treated countries (Models 3 and 4) as robustness check. Our preferred specifications (Models 1 and 2) use not-yet-treated countries as the primary control group, since future ETS entrants share more similar institutional frameworks, EU integration trajectories, and energy system characteristics with treated countries. This mitigates the risk that observed emission trends in control countries reflect structural divergence rather than a counterfactual.

4. Data

We compiled a balanced panel dataset covering 1990–2022. CO_2 emissions for the treatment group—comprising the EU-27 countries, Iceland, and Norway—were obtained from the European Environment Agency (2025), specifically from national emissions reported to the United Nations Framework Convention on Climate Change and to the EU under the Governance Regulation. These data capture energy-related sectors, including public electricity and heat production (1A1a), petroleum refining (1A1b), manufacture of solid fuels and other energy industries (1A1c).

For the control group—Albania, Bosnia and Herzegovina, Georgia, Moldova, Montenegro, North Macedonia, and Serbia—and the United Kingdom, CO_2 emissions data were drawn from Climate Watch (2025) Historical GHG Emissions (1990–2022). These emissions correspond to electricity and heat production (1A1a, 1A1b and 1A1c).

Notably, Liechtenstein was excluded from the sample despite its participation in the EU ETS because data availability was limited and its emissions contribution was relatively negligible. Further, the UK was excluded after Phase 3 because of its withdrawal from the EU and the

establishment of its independent emissions trading system.

Finally, national energy balance data (1990–2022) for energy primary production by source were retrieved from Eurostat, European Commission (2025)'s Complete Energy Balances. In all cases, emission data were expressed in MtCO2e (metric tons of carbon dioxide equivalent), which is a standardized measure, and the energy production of each source in kilotons of oil equivalent (Ktoe), which is useful for comparing different energy sources by converting them into the equivalent energy content of crude oil. Data on domestic energy production by source for Bosnia (1990–2013), Georgia (1990–2012), and Moldova (1990–2009) were retrieved from the International Energy Agency (IEA) (2025). The data were originally reported in terajoules and were subsequently converted to kilotons of oil equivalent (*ktoe*). It is important to note that Serbia included Montenegro until 2005. Therefore, data specific to Montenegro are available only from 2005 onward, because it was previously part of Serbia.

Before presenting the policy evaluation results, it is useful to examine descriptive statistics for the key variables of interest. Figure 1 illustrates the evolution of emissions across all countries included in this study from 1990 to 2022. This figure indicates that a small number of countries account for a disproportionately large share of total emissions; most countries contribute relatively little. Figure 2 further breaks down the trends by group, classifying average emissions of countries based on their entry year into the EU ETS, 2005 (the majority), 2008, and 2013 (Croatia). Notably, the 2005 group consists of the largest emitters in the energy production sector, followed by the 2008 cohort. This figure highlights a pronounced downward trend in emissions for the 2005 group, a more moderate decline for the 2008 group, and a relatively stable trajectory for the 2013 group. By contrast, the control group exhibits almost no discernible trend. These patterns support the parallel trends assumption in the prepolicy period because the treated and control groups followed similar trajectories before the EU ETS introduction. Additionally, the postpolicy decline in emissions among treated countries further reinforces the policy's potential impact. This pattern is further supported by Table 2, which reports summary statistics, including the mean, standard deviation, minimum, and maximum values of CO_2 emissions in $MtCO_2$ for treated countries across different years.

Table 2—Descriptive Statistics by Year in $MtCO_2$ from 1990 to 2022

Year	Mean	Standard Deviation	Minimum	Maximum
1990	49.41	86.52	.013	427.95
1991	48.10	84.30	.015	414.24
1992	46.65	81.37	.014	392.69
1993	44.40	77.59	.014	381.69
1994	44.11	77.01	.014	379.23

1995	44.23	75.12	.015	367.49
1996	45.16	76.49	.012	374.34
1997	44.00	72.90	.007	353.34
1998	44.29	73.24	.009	356.25
1999	42.92	71.53	.007	344.19
2000	44.08	74.35	.006	357.13
2001	45.04	76.23	.006	369.66
2002	45.57	76.49	.007	371.04
2003	47.36	79.59	.005	387.49
2004	47.12	79.55	.003	385.28
2005	45.64	78.46	.003	378.07
2006	46.08	79.19	.009	380.51
2007	46.13	79.20	.025	384.98
2008	44.05	75.54	.01	366.75
2009	40.67	69.20	.008	342.68
2010	41.28	71.09	.008	353.99
2011	40.81	69.81	.006	349.41
2012	40.62	71.83	.008	360.35
2013	38.51	70.64	.004	364.73
2014	35.98	66.11	.005	345.67
2015	35.80	63.98	.004	333.56
2016	34.51	62.06	.002	329.83
2017	34.14	59.33	.002	310.29
2018	32.27	56.81	.002	297.69
2019	28.80	48.93	.005	247.48
2020	23.39	41.71	.003	209.38
2021	25.01	47.01	.003	236.60
2022	25.73	48.33	.01	247.45

Consistent with the earlier findings, the data reveal a clear downward trend in emissions over time. However, the substantial variation in emissions, as indicated by the high standard deviation and wide range between minimum and maximum values, highlights significant heterogeneity among countries, likely driven by differences in economic size and industrial activity. Table 3 further examines emissions at the country level, highlighting stark differences between large EU economies and smaller non-ETS nations.

Table 3—Descriptive Statistics by Country in $MtCO_2$ from 1990 to 2022

Year	Mean	Standard Deviation	Minimum	Maximum
Albania	0.28	0.16	0.05	0.73
Austria	12.34	2.11	8.36	15.95
Belgium	25.09	4.13	18.06	30.24
Bosnia and Herzegovina	10.81	4.29	0.69	16.58
Bulgaria	27.71	3.68	17.82	36.40
Croatia	5.64	1.17	3.66	7.87
Cyprus	3.01	0.61	1.76	3.99
Czechia	57.08	6.58	40.78	66.21

Denmark	23.37	9.57	7.18	44.48
Estonia	13.63	4.40	5.72	28.27
Finland	23.03	6.11	12.65	37.12
France	55.60	9.12	36.97	68.94
Georgia	2.88	3.67	0.87	17.63
Germany	348.53	50.19	209.39	427.95
Greece	46.58	9.84	24.43	59.37
Hungary	19.03	4.78	10.67	26.62
Iceland	0.01	0.01	0.00	0.03
Ireland	13.03	2.23	8.51	17.24
Italy	128.18	23.65	81.21	161.49
Latvia	2.58	1.24	0.96	6.30
Lithuania	5.39	2.79	2.22	14.59
Luxembourg	0.52	0.47	0.03	1.30
Malta	1.57	0.50	0.58	2.16
Moldova	5.10	3.00	2.81	13.66
Montenegro	1.41	0.25	0.82	1.77
Netherlands	61.56	6.70	44.95	70.62
North Macedonia	5.39	1.01	3.26	6.95
Norway	12.24	2.52	7.12	15.21
Poland	178.80	21.47	138.99	234.29
Portugal	18.06	4.28	8.20	25.38
Romania	41.15	14.84	18.04	71.45
Serbia	33.18	2.85	26.44	40.41
Slovakia	10.82	2.85	6.37	19.01
Slovenia	5.60	0.83	3.40	6.70
Spain	85.67	21.69	40.34	125.54
Sweden	10.25	1.60	7.35	15.53
United Kingdom	206.02	44.18	97.88	248.62

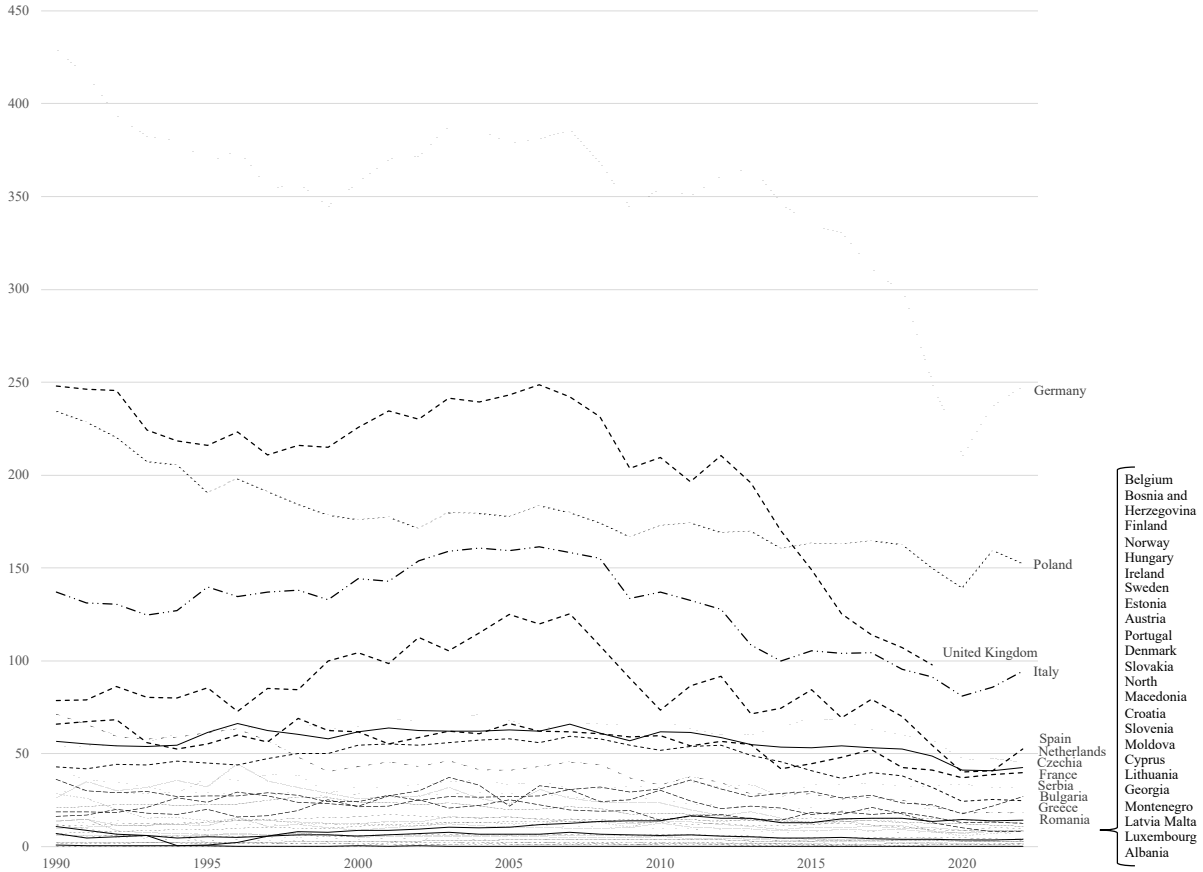

Figure 1: CO_2 emissions in $MtCO_2$ during 1990–2020 by country.

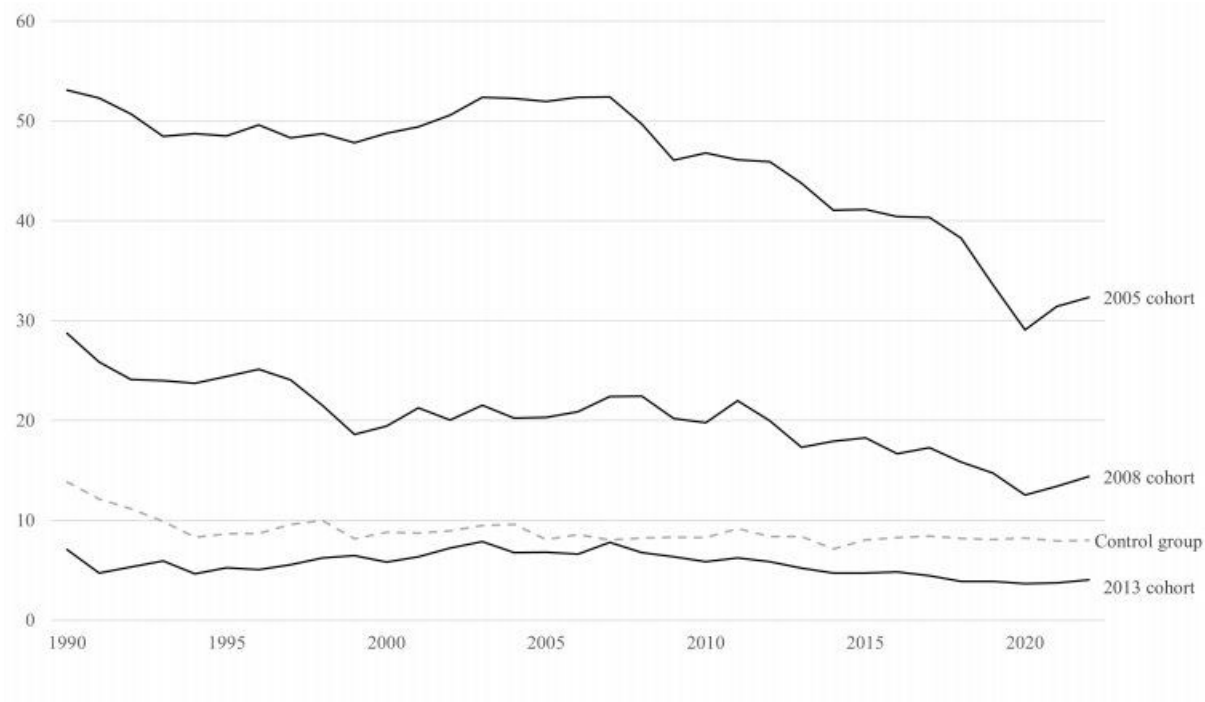

Figure 2. Average CO_2 emissions in $MtCO_2$ during 1990–2020 by cohort

Unsurprisingly, major industrial economies such as Germany, France, Poland, and Italy dominate in absolute emission levels. However, these countries have also achieved significant reductions over time, in contrast to control countries such as Albania and Montenegro, which contribute marginally to total emissions and exhibit minimal changes.

5. Results

Treatment effects were estimated with a DiD design that recognizes the staggered implementation of the EU ETS across member states. To ensure reliable inference, standard errors are clustered at the country level, thereby addressing potential serial correlation within each national panel over time. This procedure follows established practice in the STATA's *csdid* package through the ivar(id) specification to declare the country identifier for clustering.

The decision to employ one-way clustering at the country level rather than two-way clustering was motivated by several important considerations. First, treatment cohorts exhibited relatively limited variation, comprising only four distinct groups (including the never-treated units). This modest number of clusters falls short of the conventional threshold recommended in the econometrics literature for reliable cluster-robust inference. As emphasized by Cameron and Miller (2015), cluster-robust variance estimation typically requires a sufficiently large number of clusters to maintain proper size control of hypothesis tests. Second, as demonstrated by Angrist and Pischke (2008) through extensive simulations, clustering along dimensions with few groups can lead to severe underestimation of standard errors, resulting in inflated Type I error rates. Restricting clustering to the country level leads to a more conservative approach to inference that respects the panel structure of the data while avoiding potential over-rejection of null hypotheses. This methodological choice aligns with established best practices in DiD applications with staggered treatment timing, as recommended by Callaway and Sant'Anna (2021). The results of the estimation for each cohort and group can be found in Table 4.

Table 4—EU ETS Estimation Results

	Model (1) $MtCO_2$	Model (2) $MtCO_2$	Model (3) $MtCO_2$	Model (4) $MtCO_2$
				g2005
t_2004_2005	-0.121	-0.205	0.0533	-0.0521
	(0.792)	(0.769)	(0.872)	(0.816)
t_2004_2006	-0.194	-0.154	-0.0909	-0.107
	(0.663)	(0.645)	(0.699)	(0.679)

t_2004_2007	-0.727	-0.607	0.241	0.114
	(0.871)	(0.733)	(1.016)	(0.890)
t_2004_2008	-2.829**	-2.919***	-2.800**	-2.878**
	(1.183)	(1.020)	(1.287)	(1.125)
t_2004_2009	-7.700***	-7.740***	-7.772***	-7.800***
	(2.420)	(2.355)	(2.460)	(2.395)
t_2004_2010	-6.334***	-6.411***	-6.430***	-6.487***
	(2.421)	(2.337)	(2.466)	(2.382)
t_2004_2011	-8.662***	-8.504***	-8.863***	-8.716***
	(2.531)	(2.520)	(2.577)	(2.567)
t_2004_2012	-7.690***	-7.728***	-7.837***	-7.858***
	(2.112)	(2.007)	(2.180)	(2.078)
t_2004_2013	-10.34***	-10.27***	-10.34***	-10.27***
	(3.004)	(2.979)	(3.004)	(2.979)
t_2004_2014	-12.41***	-12.67***	-12.41***	-12.67***
	(3.987)	(3.755)	(3.987)	(3.755)
t_2004_2015	-14.07***	-14.14***	-14.07***	-14.14***
	(4.277)	(4.222)	(4.277)	(4.222)
t_2004_2016	-16.07***	-16.07***	-16.07***	-16.07***
	(5.217)	(5.184)	(5.217)	(5.184)
t_2004_2017	-16.92***	-16.86***	-16.92***	-16.86***
	(5.753)	(5.759)	(5.753)	(5.759)
t_2004_2018	-19.00***	-19.02***	-19.00***	-19.02***
	(6.301)	(6.263)	(6.301)	(6.263)
t_2004_2019	-23.91***	-23.96***	-23.91***	-23.96***
	(7.643)	(7.599)	(7.643)	(7.599)
t_2004_2020	-24.12***	-24.16***	-24.12***	-24.16***
	(7.777)	(7.742)	(7.777)	(7.742)
t_2004_2021	-21.27***	-21.43***	-21.27***	-21.43***
	(6.797)	(6.713)	(6.797)	(6.713)
t_2004_2022	-20.54***	-20.63***	-20.54***	-20.63***
	(6.180)	(6.112)	(6.180)	(6.112)
				g2008
t_2007_2008	0.0940	0.0374	-0.0441	-0.132
	(0.570)	(0.589)	(0.564)	(0.585)
t_2007_2009	-2.208	-2.329	-2.416	-2.595
	(1.886)	(1.973)	(1.881)	(1.978)
t_2007_2010	-2.447	-2.518	-2.699	-2.820
	(2.807)	(2.846)	(2.801)	(2.847)
t_2007_2011	-1.128	-1.476	-1.452	-1.923
	(2.452)	(2.573)	(2.455)	(2.582)
t_2007_2012	-2.489	-2.612	-2.777	-2.968
	(2.446)	(2.531)	(2.438)	(2.538)
t_2007_2013	-5.301	-5.649	-5.301	-5.649
	(3.599)	(3.824)	(3.599)	(3.824)
t_2007_2014	-3.461	-3.257	-3.461	-3.257
	(3.739)	(3.692)	(3.739)	(3.692)
t_2007_2015	-4.077	-4.165	-4.077	-4.165
	(3.845)	(3.893)	(3.845)	(3.893)
t_2007_2016	-5.861	-6.083	-5.861	-6.083
	(4.224)	(4.374)	(4.224)	(4.374)
t_2007_2017	-5.407	-5.745	-5.407	-5.745
	(4.128)	(4.328)	(4.128)	(4.328)
t_2007_2018	-6.584	-6.791	-6.584	-6.791
	(4.388)	(4.534)	(4.388)	(4.534)

t_2007_2019	-7.656	-7.781	-7.656	-7.781
	(5.014)	(5.093)	(5.014)	(5.093)
t_2007_2020	-10.03*	-10.22*	-10.03*	-10.22*
	(5.728)	(5.873)	(5.728)	(5.873)
t_2007_2021	-8.730	-8.764	-8.730	-8.764
	(5.483)	(5.517)	(5.483)	(5.517)
t_2007_2022	-7.888	-7.994	-7.888	-7.994
	(5.657)	(5.723)	(5.657)	(5.723)
				g2013
t_2012_2013	-0.493*	-0.473**	-0.493*	-0.473**
	(0.295)	(0.239)	(0.295)	(0.239)
t_2012_2014	0.238	0.197	0.238	0.197
	(0.633)	(0.516)	(0.633)	(0.516)
t_2012_2015	-0.725	-0.737*	-0.725	-0.737*
	(0.444)	(0.415)	(0.444)	(0.415)
t_2012_2016	-0.798**	-0.795**	-0.798**	-0.795**
	(0.395)	(0.380)	(0.395)	(0.380)
t_2012_2017	-1.336***	-1.319***	-1.336***	-1.319***
	(0.418)	(0.378)	(0.418)	(0.378)
t_2012_2018	-1.614***	-1.611***	-1.614***	-1.611***
	(0.311)	(0.300)	(0.311)	(0.300)
t_2012_2019	-1.612***	-1.619***	-1.612***	-1.619***
	(0.287)	(0.271)	(0.287)	(0.271)
t_2012_2020	-2.002***	-2.003***	-2.002***	-2.003***
	(0.399)	(0.384)	(0.399)	(0.384)
t_2012_2021	-1.520***	-1.538***	-1.520***	-1.538***
	(0.329)	(0.286)	(0.329)	(0.286)
t_2012_2022	-1.310***	-1.318***	-1.310***	-1.318***
	(0.175)	(0.158)	(0.175)	(0.158)
Observations	1,203	1,203	1,203	1,203

Notes: Standard errors in parentheses. * $P < 0.10$, ** $P < 0.05$, *** $P < 0.01$. Models 1 and 3 do not include covariates. Models 2 and 4 include renewable share. Models 1 and 2 use not-yet-treated and never-treated countries as the control group. Models 3 and 4 use only never-treated countries.

5.1. Model 1

This model, which uses not-yet-treated and never treated countries as the control group, estimated an average treatment effect on the treated (ATT) of -10.75 $MtCO_2$—significant at the 1% level. This suggests that the EU ETS reduced annual emissions by an average of 10.75 $MtCO_2$ per country relative to the counterfactual (no policy). Given that mean emissions across treated countries in 2004—the year before Phase I implementation—stood at 54.96 $MtCO_2$, this reduction corresponds to a 19.55% decline in emissions compared with the baseline. The result captures the cumulative effect across all implementation phases (2005–2022), suggesting the EU ETS achieved substantial abatement despite heterogeneity in treatment timing and country-specific trends.

The analysis of cohort-specific effects reveals significant heterogeneity in treatment responses.

Countries that entered the EU ETS during its initial implementation phase in 2005, comprising the largest and most economically influential group, exhibited the strongest policy response. On average, these early adopters achieved a total annual reduction of 11.76 million metric tons of CO_2 across all periods, corresponding to an 18.88% decrease relative to baseline emissions among treated countries in this cohort. This outcome aligns with theoretical expectations because earlier entrants were exposed to the cap-and-trade system for a longer duration and were therefore more likely to undertake deeper structural adjustments in their energy systems and industrial production processes.

By contrast, the 2008 cohort experienced a more modest average annual reduction of 4.87 $MtCO_2$; however, this estimate is not statistically significant ($p = 0.179$). Croatia, which joined the EU ETS as a single-country cohort in 2013, recorded a smaller but statistically significant average annual reduction of 1.11 million metric tons of CO_2 but still represented an average 19% emissions reduction compared with the baseline year (2012), attributable to its participation in the emissions trading system. Although this estimate is quantitatively lower than those observed for earlier cohorts, it remains policy-relevant in the context of Croatia's smaller economic scale and later entry into the program.

The temporal evolution of treatment effects, estimated through calendar-time aggregation, offers valuable insights into the policy's impact. When evaluating the treatment effect by period aggregated across all treated countries rather than by cohort over time, the results remained consistent. Significant emission reductions emerged in 2009, coinciding with the start of Phase II of the EU ETS, which introduced stricter emissions caps and restrictions on international credit usage. The average annual treatment effect increased steadily from -6.94 $MtCO_2$ on average in 2009 to -18.12 $MtCO_2$ by 2022, reflecting the policy's growing effectiveness as it matured. The sustained increase in emission reductions suggests that the EU ETS's dynamic design reinforced long-term decarbonization in the energy sector.

The dynamic treatment effects, presented in Figure 3, provide compelling evidence for a policy-effect interpretation of these results. The absence of statistically significant pretreatment trends supports the validity of the parallel trends assumption that underlies our DiD approach. In the posttreatment period, we observed statistically significant reductions that generally increased with the duration of program exposure. This pattern is particularly evident for the 2005 cohort, in which the effects grew steadily over time, reaching their maximum impact approximately fifteen years after initial implementation.

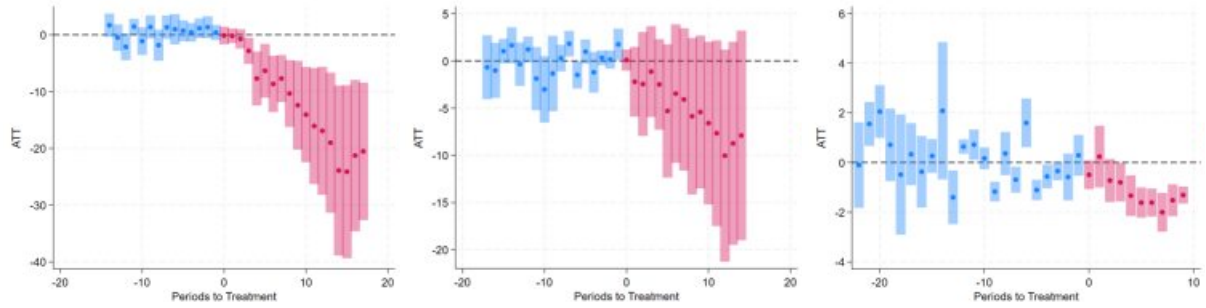

Figure 3: Dynamic ATT for 2005, 2008, and 2013 cohorts using model 1. Blue bars denote pretreatment periods; red bars denote posttreatment periods.

5.2. Model 2

To address potential systematic differences between treatment and control groups, we implemented additional specifications incorporating covariate adjustment. Model 2 introduced control for renewable energy penetration, operationalized as the ratio of renewable energy production to total primary energy production, while maintaining never-treated and not-yet-treated countries as the control group. This dual specification strategy served to both verify the robustness of the previous specification estimates and examine how treatment effects vary conditional on pretreatment energy mix characteristics.

The renewable energy share covariate served two important purposes in the analysis. First, it helped account for pre-existing differences in energy mix that might independently affect emissions trajectories. Second, it addressed potential imbalances between treatment and control groups that could bias the estimates.

Remarkably, the ATT estimates remained highly stable compared to Model 1. This second specification produced an overall average treatment effect on the treated (ATT) of -10.52 million metric tons of CO_2 per year ($P < 0.01$) across all cohorts and years, virtually identical to the estimate obtained in the baseline model. The consistency of this result across alternative specifications strengthens confidence in the reliability and validity of the estimated treatment effect.

At the cohort level, the results exhibited strong consistency with the findings from the previous model. The 2005 cohort continued to demonstrate the largest average annual reduction in emissions, at 11.79 million metric tons. The effect for the 2008 cohort remained statistically insignificant (-5.02 $MtCO_2$, $P > 0.10$) whereas Croatia's estimated reduction remained almost unchanged at 1.12 $MtCO_2$. This consistent pattern across cohorts indicates that the findings are robust to variations in the inclusion of covariates, thereby reinforcing the validity of the estimated treatment effects.

The calendar-year analysis under this specification indicates that emission reductions became

statistically significant as early as 2008, slightly preceding the point of significance in the previous model. The dynamic treatment effects, shown in Figure 4, closely mirror those observed in Model 1, reaching a peak average reduction of -24.15 $MtCO_2$ in 2020 before moderating slightly to -20.63 $MtCO_2$ by 2022, across all groups. This trajectory may reflect both the full implementation of Phase III reforms and emerging saturation in the potential for further emission reductions.

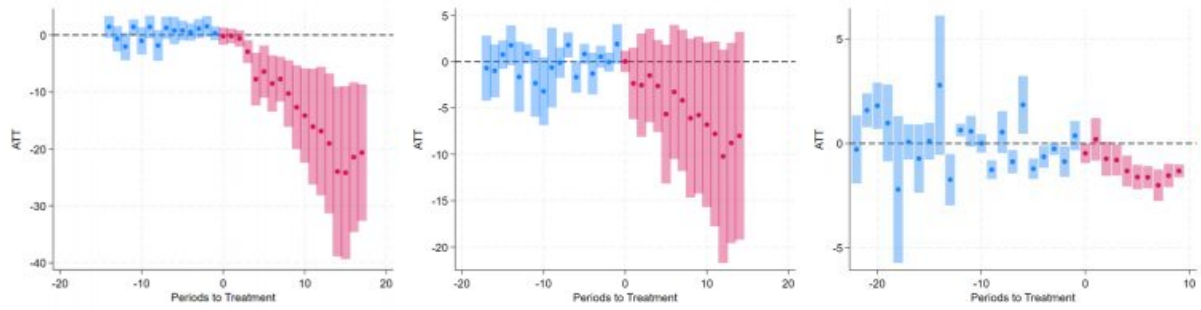

Figure 4: Dynamic ATT for 2005, 2008, and 2013 cohorts using model 2. Blue bars denote pretreatment periods; red bars denote posttreatment periods.

5.3. Models 3 and 4

Models 3 and 4 follow the same specification and covariate structure as Models 1 and 2, respectively, with the sole distinction that the control group is restricted to never-treated countries. Remarkably, the ATT estimates remained highly stable across these specifications: a reduction of 10.45 $MtCO_2$ in Model 3 and 10.78 $MtCO_2$ in Model 4 on average across cohort-time groups, both significant at the 1% level. This consistency suggests that the Model 1 results were not substantially confounded by differences in control group composition. The cohort-level effects maintained their familiar pattern across Models 3 and 4. The 2005 cohort showed average annual reductions of 11.71 $MtCO_2$ (Model 3) and 11.76 $MtCO_2$ (Model 4) whereas Croatia's effect remained at approximately -1.11 $MtCO_2$. The 2008 cohort continued to show statistically insignificant effects in both specifications. The dynamic treatment effects, presented in Figures 5 and 6, revealed several notable patterns. First, the policy's impact intensified over time across all entry cohorts relative to the 2005 baseline. Second, the effect peaked around 2020, with average annual emission reductions reaching 24.12 million metric tons of CO_2. Third, a slight attenuation was observable by 2022 because reductions declined to 20.53 $MtCO_2$ although they remained statistically significant. Significantly, no substantial differences emerged between Model 3 and Model 4, both of which yielded closely aligned results—further reinforcing the robustness and consistency of the estimated effects.

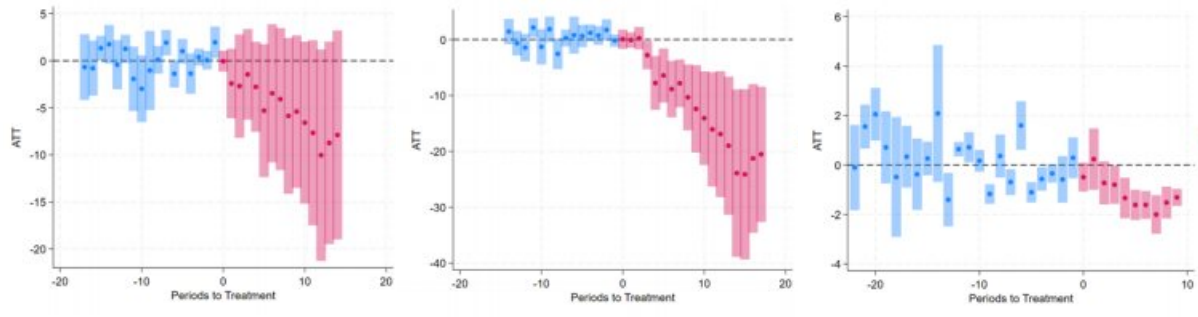

Figure 5. Dynamic ATT for 2005, 2008, and 2013 cohorts using model 3. Blue bars denote pretreatment periods; red bars denote posttreatment periods.

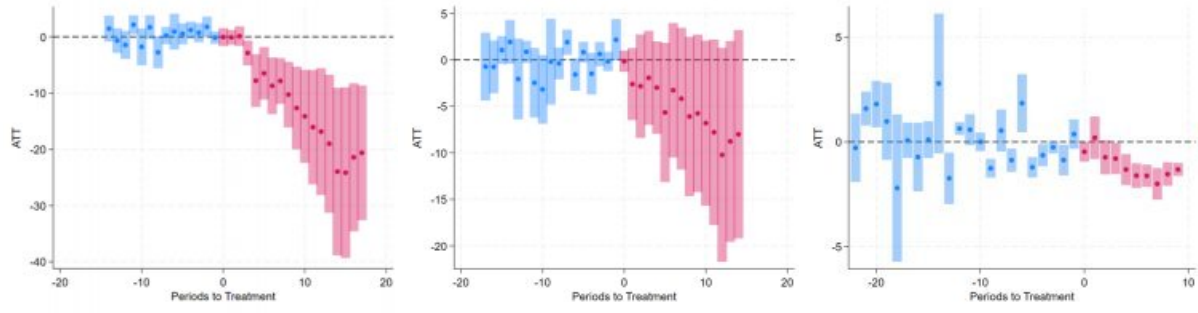

Figure 6. Dynamic ATT for 2005, 2008, and 2013 cohorts using model 4. Blue bars denote pretreatment periods; red bars denote posttreatment periods.

5.4. Reassessing the 2008 Cohort Results

To assess whether the non-significance of the 2008 cohort across all specifications stemmed from model misspecification or from high variability within the cohort, the same four model specifications were re-estimated after dividing the cohort into two subgroups: the Balkans (Romania and Bulgaria) and the Nordics (Iceland and Norway). We first excluded the Balkan countries to identify the Nordic effect and subsequently excluded the Nordics to estimate the Balkan effects. In both cases, the excluded subgroup remained part of the control pool for the 2005 cohort, thereby minimizing the loss of observations. The subgroup analysis of the 2008 cohort revealed that the initial lack of statistical significance in the estimates masked substantial heterogeneity between Nordic and Balkan countries.

Across all four model specifications, the Nordic subgroup (Iceland and Norway) showed either negligible or slightly positive treatment effects, indicating that the EU ETS had no discernible impact on emission reductions in these countries and may even have coincided with moderate increases. In the baseline model without covariates (Model 1), the average treatment effect on the 2008 cohort shifted from -4.87 $MtCO_2$ (insignificant) in the combined sample to

0.38 $MtCO_2$, still not significant at the 10% level when only the Nordic countries were considered. This result remained stable across Models 2 to 4, where the ATT fluctuated narrowly around 0.3 $MtCO_2$ per year and remained significant at conventional confidence levels. The persistence of this positive and significant effect, even after accounting for additional covariates and alternative control groups, suggests that the Nordic emissions trajectory diverged structurally from that of other EU ETS participants. This likely reflects their already low carbon intensity, high renewable energy penetration, and limited exposure to fossil-based generation within the trading system.

By contrast, the Balkan subgroup (Romania and Bulgaria) displayed a different response to the EU ETS, opposite to that of the Nordics but consistent with the pattern observed for the 2005 cohort. Once isolated from the Nordic countries, the estimated ATT became strongly negative and statistically significant across all models. In Model 1, the ATT for the 2008 subgroup shifted from an annual average of -4.87 $MtCO_2$ (insignificant) to -10.14 $MtCO_2$, significant at the 95% level. This improvement remained consistent when including not-yet-treated countries as controls (Model 2) and when introducing the renewable energy share as a covariate under the double-robust estimation framework (Models 3 and 4). This evidence indicates that the EU ETS was substantially more effective in the Balkans, likely due to higher baseline emissions, greater reliance on fossil fuels, and stronger marginal incentives for abatement once the system was implemented.

Overall, these results demonstrate that the lack of significance in the baseline 2008 results was driven by high within-cohort heterogeneity rather than model misspecification. When the 2008 cohort was disaggregated, both subgroups showed statistically significant effects, but in opposite directions. Dynamic treatment-effect plots confirm this divergence (see Figure 7), showing a persistent upward trajectory for the Nordics and a sustained downward trend for the Balkans after 2013. These contrasting dynamics are also descriptively visible in Figure 8. The results of the estimation can be found in Table 5.

Table 5—EU ETS Estimation Results for 2008 Cohort Divided into Subgroups

	Model (1) Nordic subgroup $MtCO_2$	Model (1) Balkan subgroup $MtCO_2$	Model (2) Nordic Subgroup $MtCO_2$	Model (2) Balkan subgroup $MtCO_2$
		g2008		
t_2007_2008	0.0798	0.108	0.123	-0.254
	(0.225)	(1.071)	(0.261)	(1.077)
t_2007_2009	0.465	-4.880*	0.545**	-5.762**
	(0.466)	(2.584)	(0.252)	(2.519)
t_2007_2010	0.751	-5.646	0.801**	-6.269
	(0.540)	(4.557)	(0.371)	(4.499)
t_2007_2011	-0.333	-1.924	-0.0869	-3.932
	(0.714)	(4.695)	(0.860)	(4.566)

t_2007_2012	0.161	-5.138	0.243	-6.150
	(0.458)	(4.041)	(0.378)	(3.982)
t_2007_2013	0.0170	-10.62**	0.190	-11.80**
	(0.576)	(4.754)	(0.848)	(4.717)
t_2007_2014	1.586**	-8.508	1.496	-7.795
	(0.773)	(5.386)	(1.043)	(5.511)
t_2007_2015	0.893	-9.047	0.938*	-9.364
	(0.730)	(5.792)	(0.492)	(5.795)
t_2007_2016	0.465	-12.19**	0.573	-13.12**
	(0.738)	(5.487)	(0.490)	(5.470)
t_2007_2017	0.480	-11.29**	0.647	-12.52**
	(0.840)	(5.657)	(0.562)	(5.626)
t_2007_2018	0.705	-13.87***	0.809*	-14.70***
	(0.771)	(4.758)	(0.445)	(4.755)
t_2007_2019	0.662	-15.97***	0.726**	-16.42***
	(0.592)	(5.546)	(0.311)	(5.529)
t_2007_2020	0.138	-20.20***	0.228	-20.97***
	(0.609)	(5.173)	(0.531)	(5.164)
t_2007_2021	0.0661	-17.53***	0.0842	-17.80***
	(0.490)	(6.495)	(0.473)	(6.485)
t_2007_2022	-0.332	-15.44*	-0.277	-15.86*
	(0.491)	(8.392)	(0.696)	(8.371)
Observations	1,173	1,173	1,173	1,173

Notes: Standard errors in parentheses. *$P < 0.10$, **$P < 0.05$, ***$P < 0.01$. Model 1 does not include covariates. Model 2 includes renewable share. Models 1 and 2 use not-yet-treated and never-treated countries as the control group.

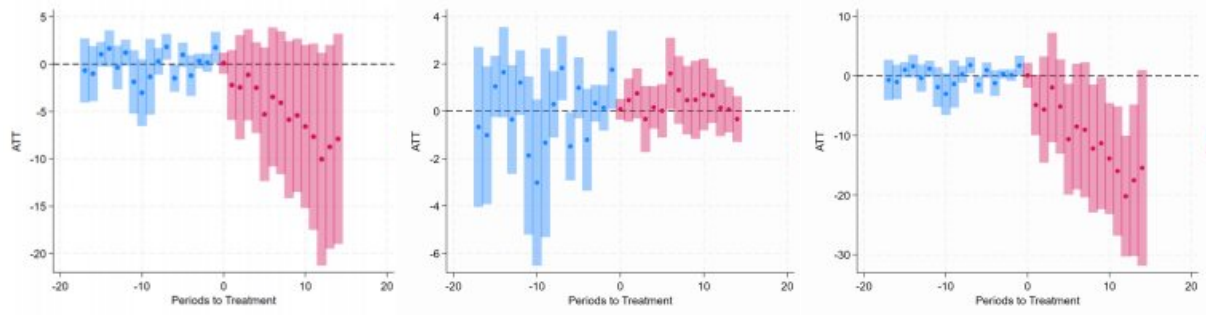

Figure 7. Dynamic ATT estimates for the 2008 cohort, shown jointly and by subgroups using model 1. Blue bars denote pretreatment periods; red bars denote posttreatment periods. from left to right: (i) full sample, (ii) Norway and Iceland, and (iii) Romania and Bulgaria.

The parallel trends assumption was not retested for the subgroups because observations were only excluded after 2008; therefore, the pretrends robustness checks and results remained unaffected by dividing the 2008 cohort. However, linearity, covariate balance, endogeneity, and overlap were assessed for the subgroups individually, and no significant differences were found compared to the results for the combined cohort. This provides evidence that the estimation results derived from the 2008 subgroups are robust and consistent with the overall analysis.

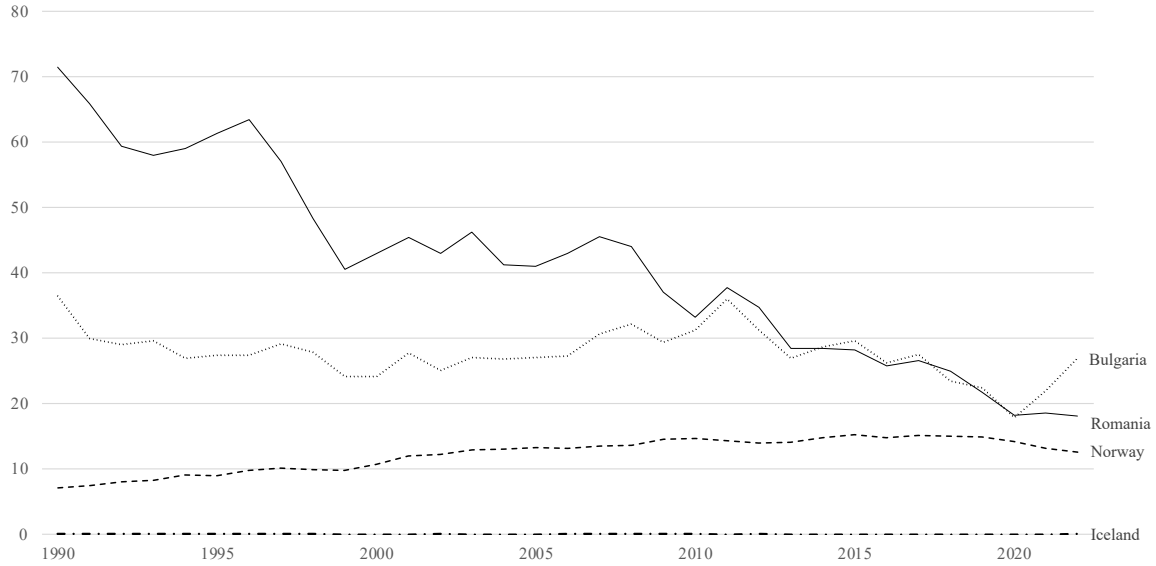

Figure 8. CO_2 emissions in $MtCO_2$ from 1990–2020 for 2008 cohort

5.5. Comparison with TWEE and Goodman-Bacon Decomposition

To evaluate how the choice of estimator influences the results, we estimated a conventional two-way fixed effects (TWFE) model using the same sample as in the staggered DiD analysis. We then applied the Goodman-Bacon (2021) decomposition to identify the sources of variation underlying the TWFE estimate and assess the extent to which it relies on potentially contaminated comparisons between treatment cohorts.

Table 6 reports the TWFE estimates for the full sample and by treatment cohort, with and without renewable energy share as a covariate. For the full sample, the estimated ATT equals −4.80 $MtCO_2$ ($p < 0.10$) in the baseline specification and −2.27 $MtCO_2$ when renewable energy share is included, the latter becoming statistically insignificant. These estimates are substantially smaller in magnitude than the ATT estimates obtained using the Callaway and Sant'Anna (2021) estimator, which range between −10.45 and −10.78 $MtCO_2$ across the four main specifications. The attenuation of the TWFE coefficient is consistent with the presence of treatment-effect heterogeneity and dynamic treatment effects, both of which are evident in the event-study results presented in Sections 5.1–5.3. As treatment effects increase with exposure duration, TWFE mechanically averages effects across cohorts and periods, potentially biasing the aggregate estimate toward zero.

At the cohort level, the TWFE estimates broadly reproduce the direction of the effects identified by the staggered DiD approach but generally with smaller magnitudes and lower

statistical precision. The 2005 cohort shows a significant reduction of -8.74 $MtCO_2$ in the baseline specification, compared with an average ATT of approximately -11.76 $MtCO_2$ under the Callaway and Sant'Anna (2021) framework. For the 2008 cohort, the estimated effect remains statistically insignificant, mirroring the results obtained in the main analysis. Similarly, the TWFE estimate for Croatia's 2013 cohort is negative but statistically insignificant, whereas the staggered DiD estimator identifies a significant reduction in emissions. The subgroup analysis of the 2008 cohort produces the same qualitative pattern observed previously: positive coefficients for the Nordic countries and negative coefficients for the Balkan countries, reflecting substantial within-cohort heterogeneity.

The Goodman-Bacon decomposition provides further insight into these differences. In the full sample, treated-versus-never-treated comparisons account for approximately 77% of the total weight and generate a coefficient of -6.72 $MtCO_2$. These comparisons represent the cleanest source of identification because they do not rely on already-treated units serving as controls. The remaining 23% of the weight is assigned to comparisons between treatment cohorts at different stages of exposure. These Early-versus-Late and Late-versus-Early comparisons generate substantially weaker and, in some cases, positive estimates, thereby pulling the overall TWFE coefficient toward zero. The decomposition therefore confirms that the aggregate TWFE estimate is influenced by comparisons that are vulnerable to bias when treatment effects vary across cohorts and over time.

Table 6—TWFE Estimates: Overall and by Cohort.

	(1) Overall	(2) Overall (cov)	(3) 2005	(4) 2005 (cov)	(5) 2008	(6) 2008 (cov)	(7) 2013	(8) 2013 (cov)
Treated	-4.802*	-2.272	-8.742***	-5.854*	-4.550	-3.105	-1.310	0.277
	(2.526)	(2.485)	(3.183)	(3.002)	(6.065)	(5.332)	(1.201)	(1.105)
Renewable share		-18.42		-15.85		-22.49		-8.257
		(12.52)		(12.15)		(12.81)		(6.071)
Observations	1203	1203	1038	1038	348	348	249	249

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses. Control group: never-treated countries. All models include country and year fixed effects. Standard errors clustered at country level.

Table 7—TWFE Estimates: 2008 Cohort Full Sample and Subgroups.

	(1) 2008 Full	(2) 2008 Full (cov)	(3) Nordic	(4) Nordic (cov)	(5) Balkan	(6) Balkan (cov)
Treated	-4.550	-3.105	2.393	2.312	-11.49	-9.813
	(6.065)	(5.332)	(2.074)	(2.107)	(9.072)	(8.997)

Renewable share		-22.49		-12.98		-12.47*
		(12.81)		(7.177)		(6.246)
Observations	348	348	318	318	318	318

Notes: Standard errors in parentheses. Control group: never-treated countries. Nordic subgroup: Iceland and Norway. Balkan subgroup: Romania and Bulgaria. Balkans/Nordics excluded post-2008 in respective subgroup.

6. Robustness Checks

This section assesses the robustness of the main findings through a comprehensive set of checks, including diagnostic tests for multicollinearity, endogeneity, linearity, covariate balance, propensity score overlap, and parallel trends; a TWFE benchmark with Goodman-Bacon decomposition; a log(CO_2) specification; the Sun and Abraham (2021) interaction-weighted estimator; the synthetic difference-in-differences estimator of Arkhangelsky et al. (2021); wild cluster bootstrap inference; a leave-one-out sensitivity analysis over the control group; a sample restricted to 1990–2019 to rule out COVID-19 confounding; and an estimation excluding the 2008–2010 Global Financial Crisis years. Across all checks, the main findings remain stable.

Diagnostic Tests and Assumption Validation

Clustered standard errors effectively addressed concerns regarding heteroskedasticity and autocorrelation whereas below 2.5 indicated no substantive multicollinearity issues because values exceeding five or ten are typically viewed as problematic, as established in the literature (Fox and Weisberg 2018; James et al. 2021). Placebo tests implementing artificial treatment assignments in preperiod periods consistently yielded statistically insignificant effects ($P > 0.10$) when conducted both individually and jointly with the renewable share covariate, providing evidence against estimates being biased by endogeneity. Last, OLS assumes a linear relationship between the covariates and the outcome. This was assessed by adding quadratic terms of the covariates in panel regressions. The results indicate that linearity cannot be rejected. In conclusion, the diagnostic tests conducted for the inclusion of covariates in the outcome regression model indicated that the assumptions required for valid inference were largely satisfied.

To test covariate balance, inverse probability weighting (IPW) was applied separately for the 2005, 2008, and 2013 treatment cohorts using a logit model based on pretreatment data to estimate the probability of treatment as a function of renewable energy share. We then used weighted regressions and linear combination tests to evaluate differences in covariate means between treated

and control units. The results indicated no significant imbalance for the 2005 and 2008 cohorts, confirming comparability between treated and control groups. Although the 2013 cohort exhibited some imbalance, this did not materially affect the overall ATT estimates, as demonstrated by their consistency across alternative specifications.

A natural concern with the propensity-score specification is that the estimated scores are concentrated within a narrow range (approximately 0.82 to 0.845), reflecting the limited ability of renewable energy share to distinguish treated from control units. The pseudo-R^2 of the cohort-specific logit models is correspondingly low (below 0.05 in all cohorts). Rather than undermining the analysis, this feature explains why the estimates are so stable across specifications: because the propensity scores vary little across units, the inverse-probability weighting applied in Models 2 and 4 barely reweights the sample, and the resulting ATT estimates are almost identical to the unweighted specifications in Models 1 and 3. The near-equivalence of weighted and unweighted estimates therefore indicates that our results do not depend on the selection model, and that the doubly robust estimator is not relying on a fragile first-stage specification to generate the headline findings.

The critical-parallel-trends assumption was validated through formal statistical testing using the Nguyen (2020) approach, which failed to reject the null hypothesis of parallel pretreatment trends ($P > 0.05$). This finding hold despite occasional visual suggestions of divergence when covariates are included, highlighting the importance of complementing graphical analysis with rigorous hypothesis testing. As observed, the ATT estimates remained remarkably stable across all four model variants. This consistency across alternative approaches strongly reinforces the reliability of the study. Further tests validating the underlying assumptions of the model are presented in Appendix I.

It is notable that the same assumptions underlying the main analysis were tested separately for the 2008 subgroups where appropriate. The parallel trends assumption was not retested for the subgroups because observations were only excluded after 2008; therefore, the pretrends robustness checks and results remained unaffected by dividing the 2008 cohort. However, linearity, covariate balance, endogeneity, and overlap were assessed for the subgroups individually; and no significant differences were found compared to the results for the combined cohort. This provides evidence that the estimation results derived from the 2008 subgroups are robust and consistent with the overall analysis.

Logarithmic Specification

In order to improve the robustness of the results, we re-estimate all specifications using the natural

logarithm of CO_2 emissions, allowing effects to be interpreted as percentage changes and reducing sensitivity to large emitters. The results closely match the baseline findings. The overall ATT ranges from -0.218 to -0.238 log points (≈19.6%–21.2% reduction), consistent with the 17.5%–19.6% effect in levels. The 2005 cohort shows effects of -0.197 to -0.217 log points (≈18%–20%), increasing to about -0.49 to -0.52 by 2019–2020 ($P < 0.01$), mirroring the compounding pattern in the main results. The 2008 cohort remains insignificant, while the 2013 cohort shows stable and significant reductions of about -0.205 to -0.207 log points (≈19%), closely matching the baseline estimate. Overall, the log specification confirms that the main results are robust to functional form, the full estimation results can be found in Appendix II

Sun and Abraham (2021) Specification

As a further robustness check, we re-estimate the event study using the interaction-weighted (IW) estimator of Sun and Abraham (2021), which provides an alternative approach to addressing two-way fixed effects bias while allowing for treatment effect heterogeneity. The pre-treatment coefficients are close to zero and statistically insignificant from $t = -2$ onward, consistent with the parallel trends assumption. In the post-treatment period, the IW estimates closely replicate the main dynamic pattern: effects are initially small and insignificant, then increase over time and become statistically significant from $t = +3$, reaching -17.61 $MtCO_2$ by $t = +13$ ($P < 0.01$) and -25.26 $MtCO_2$ by $t = +15$ ($P < 0.01$). The implied average post-treatment effect of -11.35 $MtCO_2$ is very close to the baseline ATT of -10.75 $MtCO_2$. Overall, the IW results confirm that both the timing and magnitude of the estimated effects are robust to the choice of estimator.

Synthetic Difference-in-Differences

As a final and more demanding robustness check, and to address the comparability of the control group, we re-estimate the treatment effects using the synthetic difference-in-differences (SDID) estimator of Arkhangelsky et al. (2021). Unlike conventional difference-in-differences, SDID constructs data-driven unit and time weights so that the weighted control units reproduce the pre-treatment trajectory of the treated units before estimation. This directly addresses the concern that never-treated countries differ structurally from ETS participants: rather than assuming parallel trends across heterogeneous units, the estimator selects and weights controls to match pre-treatment outcomes, and any remaining level differences are absorbed. We apply the estimator separately to each adoption cohort and aggregate using cohort sizes, with inference based on the jackknife.

The SDID results, summarized in Figure 9, reinforce the main findings. For the 2005 cohort, the estimated effect is approximately -9.9 $MtCO_2$ per year (statistically significant at conventional

levels), falling between the two-way fixed effects estimate (−8.7) and the Callaway and Sant'Anna (2021) estimate (−11.8). The 2008 cohort remains negative but statistically insignificant (approximately −4.3 $MtCO_2$), consistent with the main analysis. It is worth noting that only the point estimate for the single-country 2013 cohort (Croatia) is provided, given that jackknife requires at least two treated units., The estimation yields a small estimate that should be interpreted with caution. The cohort-size-weighted overall effect of approximately −8.8 $MtCO_2$ is close to the baseline ATT and well above the attenuated TWFE estimate. Because SDID matches pre-treatment trajectories by construction, the consistency of these estimates with our main specifications provides direct evidence that the headline results are not an artifact of structural differences between treated and control countries.

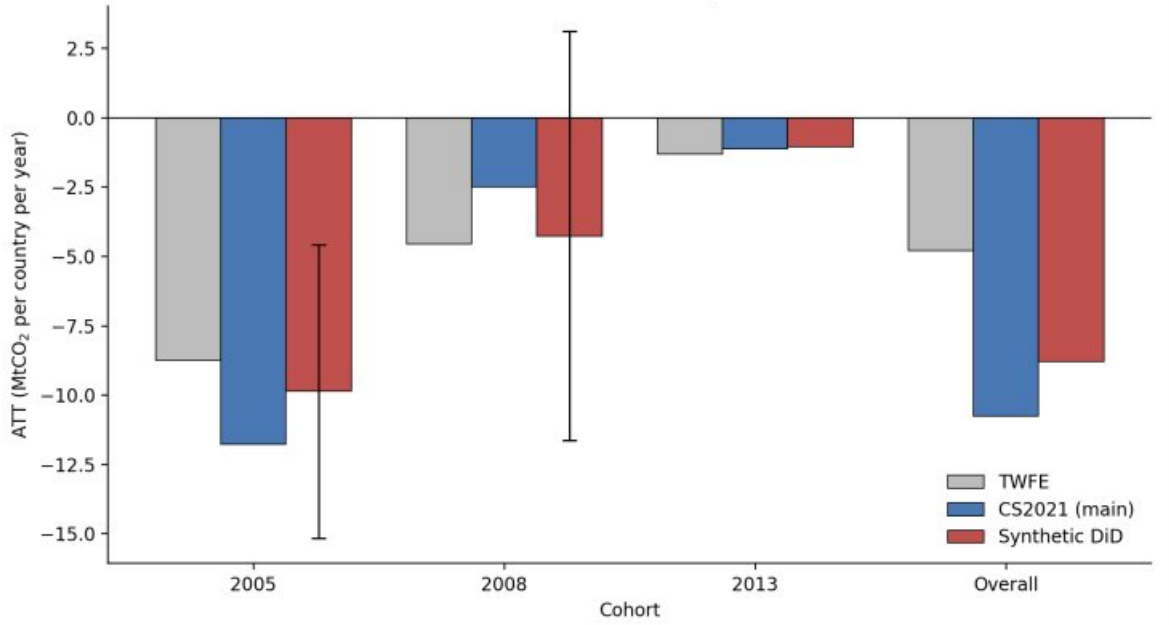

Figure 9. EU ETS effect by cohort under alternative estimators (TWFE, Callaway and Sant'Anna 2021, and synthetic DiD). Error bars show 95% confidence intervals for the synthetic DiD estimates where available.

Wild Cluster Bootstrap

As a robustness check on statistical inference, we implement the wild cluster bootstrap following Roodman et al. (2019) to address the relatively small number of clusters in our sample (37 countries). Using 999 replications with Rademacher weights, we apply the bootstrap to cohort-specific TWFE specifications restricted to each treated cohort and the never-treated control group, thereby avoiding the contamination from already-treated units identified in the Goodman-Bacon decomposition. The results, reported in Appendix II, confirm the conclusions drawn from conventional inference. The overall ATT without covariates remains significant, the specification

with covariates remains insignificant, and the 2005 cohort continues to show a statistically significant effect. Likewise, the 2008 cohort, its Nordic and Balkan subgroups, and the 2013 cohort remain statistically insignificant under both approaches. Bootstrap p-values are generally larger than their conventional counterparts, particularly for smaller cohorts, suggesting that conventional standard errors modestly understate uncertainty. Overall, the bootstrap results reinforce rather than alter the paper's main conclusions.

Leave-one-out sensitivity analysis

To ensure that the baseline results are not driven by any single control country, we conduct a leave-one-out sensitivity analysis in which each never-treated country (Albania, Bosnia and Herzegovina, Georgia, Moldova, Montenegro, North Macedonia, and Serbia) is sequentially excluded from the control group and the ATT is re-estimated. As shown in the Appendix II, the estimates remain highly stable across all iterations, closely centered around the baseline ATT of -10.76 $MtCO_2$. The point estimates range from -10.11 $MtCO_2$ (excluding Bosnia and Herzegovina) to -11.25 $MtCO_2$ (excluding Serbia), with all estimates remaining statistically significant at the 1% level and their confidence intervals fully overlapping with the baseline result.

Sensitivity to Major Economic Shocks

An additional robustness check was conducted to assess whether the large emission reductions observed in 2020 and subsequent years were driven by the COVID-19 pandemic rather than by the EU ETS itself. To this end, all four model specifications were re-estimated using a restricted sample ending in 2019. The results were virtually identical to those obtained in the full sample. The magnitude, statistical significance, and temporal evolution of the treatment effects remained unchanged across cohorts and specifications. In particular, the steadily increasing emission reductions observed for the 2005 cohort were already well established by 2019, while the 2008 cohort remained statistically insignificant and the estimated effects for Croatia's 2013 cohort were unaffected by the exclusion of the pandemic years. These findings indicate that the main results are not driven by the temporary economic disruption associated with COVID-19 and instead reflect the underlying impact of the EU ETS. Detailed results are reported in Appendix II.

Finally, to address concerns that the Global Financial Crisis (2008–2010) may confound the estimated policy effects, we re-estimate the main specification after excluding these years from the sample. The results, reported in Appendix II, are highly consistent with the baseline findings. For the 2005 cohort, effects emerge immediately after treatment and grow steadily over time, reaching approximately -23.9 $MtCO_2$ by 2019 ($p < 0.01$), closely matching the main estimates. The 2013

cohort similarly exhibits stable and statistically significant reductions of around −1.6 $MtCO_2$ in 2019, consistent with the baseline trajectory. For the 2008 cohort, estimates remain negative but less precisely estimated, reflecting the reduced post-treatment variation after excluding the crisis period. Pre-treatment coefficients remain flat and statistically insignificant across specifications.

7. Discussion

The following discussion is organized by cohort, with detailed results provided in Section 5. Starting with the 2005 cohort, the analysis indicates that across all four model specifications, this group began experiencing significant emission reductions during the second phase of the EU ETS (2008–2012). In models without covariates and using never-treated and not-yet-treated countries as controls, the first statistically significant reductions appeared around 2008 whereas other specifications detected significant impacts around 2009. This aligns with the understanding that the first phase (2005–2007) primarily served as a pilot, intended to test the operational mechanics of the program, and was characterized by design flaws such as the overallocation of allowances, leading to limited emission reductions.

By contrast, the second phase introduced critical reforms, including tighter emission caps, reduced allowances, and harsher penalties for noncompliance, which translated into meaningful emission reductions. During this period, the estimated average treatment effect ranged from approximately -2.8 to -7.8 million metric tons of CO_2 reduction per year ($P < 0.01$), depending on the model specification. These figures represent a clear departure from the largely ineffective outcomes observed during the first phase. The third implementation phase (2013–2020) saw a continuation and amplification of this downward trend. The expansion of sectoral coverage and the introduction of allowance auctioning mechanisms were associated with further emission reductions. By 2019, the estimated treatment effects relative to 2004 baseline levels reached a variation between -10.34 and -23.9 $MtCO_2$ across specifications, demonstrating the compounding impact of progressively stringent policy measures. Preliminary analysis of Phase IV (2021–2022) suggests potentially stronger treatment effects although these estimates require careful interpretation. The observed peak reduction of -24.12 $MtCO_2$ in 2020 may reflect both the policy's maturation and contemporaneous reductions in economic activity associated with the COVID-19 pandemic.

The increasing magnitude of these negative coefficients over time (from -7.7 in 2009 to -23.9 in 2019) suggests a cumulative and possibly compounding effect of the treatment for cohort 2005, in which the impact grows as time passes. Therefore, for EU–25 countries, the EU ETS effectively reduced CO_2 emissions in the energy production sector in 2022 by approximately 20 $MtCO_2$. This

is a substantial contribution to mitigating climate change, especially if it represents a single policy's impact and the effect of a specific industry. It is comparable to the annual emissions of a small country like Finland.

Turning to the 2008 cohort, which includes Romania, Bulgaria, Norway, and Iceland, the baseline results initially presented a puzzle: although the estimated average treatment effects were generally negative, they were not statistically significant. This apparent lack of significance, however, masked substantial heterogeneity within the cohort. Subgroup analysis revealed that the EU ETS had sharply divergent effects in the Balkan countries (Romania and Bulgaria) compared with the Nordic members (Iceland and Norway). Across all model specifications, the initial estimated average effect corresponded to a modest annual reduction of approximately -2.1 $MtCO_2$. This declining trend persisted throughout Phase II of the EU ETS, culminating in an estimated reduction of approximately -2.5 $MtCO_2$ by 2012, indicating a gradual yet consistent increase in the policy's mitigation impact over time, although still not statistically significant. Once this regional heterogeneity was accounted for, a clear and significant pattern emerged. The Balkan subgroup showed a strong, negative, and statistically significant response to the EU ETS, consistent with the pattern observed in the 2005 cohort. For example, in Model 1, the ATT for the 2008 cohort shifted from an insignificant -4.87 $MtCO_2$ to a significant -10.14 $MtCO_2$ on average annually. This suggests the policy was particularly effective in these countries, likely due to their higher baseline emissions and greater reliance on carbon-intensive energy. The Balkan effect became significant in 2009, immediately after implementation, at -4.88 $MtCO_2$, and peaked in 2020 with an average annual variation of -20.20 $MtCO_2$. This represents a notable decrease compared with -10.03 $MtCO_2$, which was the ATT in 2020 in the joint estimation. However, by the onset of Phase IV, the pace of additional emissions reductions appeared to decelerate. By 2022, the average reduction for the Balkan subgroup was estimated at approximately -15.44 $MtCO_2$ (7.88 $MtCO_2$ in the joint estimation). This slowdown is consistent with the pattern observed in the 2005 cohort and may reflect the disruptive impact of the COVID-19 pandemic.

By stark contrast, the Nordic subgroup shows either negligible or slightly positive treatment effects across all model specifications. When isolated, their ATT across years is an insignificant 0.3 $MtCO_2$, indicating the EU ETS had no discernible negative impact on their emissions and may even have coincided with moderate increases. Annual ATTs in the Nordic countries generally showed positive signs, but they were not statistically significant, except for 2015 and 2019, when slight increases in emissions occurred. This does not evidence of ETS failure in Iceland and Norway but rather of a ceiling effect: both countries already operated near the frontier of low-carbon electricity generation, with renewable shares consistently above 90%. Therefore, the ETS

provided little additional incentive for fuel switching because such switching had already occurred prior to participation.

By contrast, Romania and Bulgaria entered the system with high reliance on coal-fired generation and lower baseline efficiency, creating substantial marginal abatement opportunities that the carbon price signal could activate. This heterogeneity suggests that ETS effectiveness in the energy sector is closely tied to the carbon intensity of the baseline generation mix. Countries with already low-carbon systems will exhibit weaker measured effects, not because the policy fails, but because its main channel of action is largely exhausted. This has direct implications for policy design: supplementary instruments targeting sectors where price signals are insufficient, such as capacity mechanisms favoring renewables or targeted efficiency standards, may be needed to generate further reductions in already low-carbon economies.

Finally, the 2013 cohort, consisting solely of Croatia, presents a unique case within the EU ETS framework because it joined the system in the same year it entered the European Union. Despite the onset of Phase III in 2013, significant reductions in emissions did not become apparent until 2017, with an estimated reduction of -1.33 $MtCO_2$ ($P < 0.01$). This downward trend continued, reaching approximately -2 $MtCO_2$ by 2020, which indicates a gradual but consistent impact of the EU ETS on Croatia's energy sector. However, with the onset of the fourth phase, the pace of reductions appeared to slow, aligning with broader trends observed across other cohorts. By 2022, the estimated reduction in emissions stood at 1.31 $MtCO_2$. Significantly, the smaller magnitude of these coefficients was likely influenced by the country's relatively modest economic size and emissions profile, which naturally limits the potential scale of absolute reductions. This makes the estimated effects statistically significant but potentially less robust, given the lack of variability inherent in a single-country cohort, posing challenges for the precision of DiD estimates.

Overall, although the 2013 cohort shows a broadly similar downward emissions trend from earlier cohorts, the smaller scale and reduced statistical power mean that its results should be interpreted with caution. Another limitation is the low fit of the 2013 cohort in the IPW model and overlap issues, which is a crucial assumption. Nonetheless, these findings still provide valuable insights into the EU ETS's effectiveness, even for smaller economies.

Taken together, the cohort-specific results reveal that the effectiveness of the EU ETS has been shaped by both the timing of entry and the evolving design of the system. Early participants began to show meaningful reductions only after the shift from the pilot phase to the more stringent Phase II, when tighter caps and stricter compliance incentives strengthened the carbon price signal. These reforms generated progressively larger effects in subsequent years, consistent with the estimated dynamic treatment effect. Later entrants show more heterogeneous responses: the

Balkan members displayed substantial and statistically significant reductions, reflecting higher baseline carbon intensities and greater room for fuel substitution whereas the Nordic countries, already reliant on low-carbon generation, presented negligible or even slightly positive treatment effects. These patterns point to three core mechanisms: (i) carbon price levels and volatility, (ii) compliance and expectations of future stringency, and (iii) technological and fuel-switching opportunities that vary across countries. Across all cohorts, the pattern of results supports the view that carbon prices, expectations, and institutional tightening are the primary channels through which the EU ETS drives abatement. At the same time, the variability in effects indicates that the system's impact is neither uniform nor automatic.

Regarding mechanisms, although this paper provides reduced-form estimates by design, the dynamic treatment effects are consistent with several channels through which the ETS is expected to reduce energy-sector emissions. The primary mechanisms include fuel switching—substitution from coal to natural gas and renewables in electricity generation—improvements in the thermal efficiency of generation plants, and demand-side reductions in electricity consumption driven by carbon-cost pass-through to wholesale prices. The phase-specific pattern of our results aligns with these channels: the relatively weak effects during Phase I are consistent with over-allocation and near-zero prices that provided insufficient incentive for abatement, while the substantially stronger effects in Phase III coincide with tighter caps, expanded auctioning, and prices that made fuel switching from coal to gas economically viable across much of the EU generation fleet.

Several limitations of our analysis merit emphasis. The most important concerns the comparability of the control group. Any country-level evaluation of the EU ETS faces an inherent constraint: because the system covers nearly all of Europe, the only available counterfactual is a set of non-participating countries (here, the Western Balkans, Moldova, and Georgia) that differ from ETS participants in economic structure, institutional development, and pre-treatment emission levels. Unlike firm-level studies that exploit regulatory thresholds within a single jurisdiction, our setting cannot fully eliminate this concern. We address it in three ways. First, the Goodman-Bacon decomposition shows that roughly 77% of the identifying variation in the aggregate estimate, and the entirety of the 2005-cohort estimate, derives from treated-versus-never-treated comparisons rather than from contaminated comparisons between cohorts, indicating that the results are not an artifact of forbidden comparisons. Second, the leave-one-out analysis demonstrates that no single control country drives the estimates, which remain tightly centered on the baseline ATT when each control is removed in turn. Third, the absence of differential pre-treatment trends in both unconditional and conditional specifications provides empirical support for the parallel-trends assumption, although it cannot rule out unobserved time-

varying confounders. Taken together, these checks substantially mitigate, though they do not entirely eliminate, the concern that structural differences between treated and control countries bias our estimates. Accordingly, we interpret our findings as robust and policy-relevant evidence of the EU ETS's association with energy-sector emission reductions, rather than as definitive causal point estimates, and we note that any residual bias arising from control-group dissimilarity or carbon leakage to non-participating countries would most plausibly attenuate the estimated effects rather than inflate them, implying that our results may be read as conservative.

To situate our estimate within the existing literature, Table 8 compares our headline ATT with prior evaluations of the EU ETS. Our estimate of approximately a 19.6% reduction lies within the range of recent firm- and country-level studies, which report effects from roughly 3.8% (Bayer and Aklin, 2020) to 15–16% (Biancalani et al., 2024; Colmer et al., 2025). Our somewhat larger estimate is consistent with the fact that, unlike most prior work, our sample extends through Phase III and IV, when tighter caps and expanded auctioning strengthened the price signal and the associated abatement incentives.

Table 8—Comparison of estimated EU ETS emission reductions across studies. Reductions are expressed as approximate percentage declines relative to each study's baseline; methods and samples differ, so figures are not directly comparable.

Study	Method	Sample / period	Est. reduction
This paper	Staggered DiD (CS2021)	37 countries, 1990–2022, all four phases	19.6% (−10.75 $MtCO_2$)
Bayer & Aklin (2020)	Synthetic control	Sectoral, EU, 2008–2016	3.8%
Dechezleprêtre et al. (2023)	Matching + DiD	4 countries, installation-level, 2005–2012	10%
Colmer et al. (2025)	Firm-level DiD	French manufacturers, Phases I–II	14–16%
Biancalani et al. (2024)	Matrix completion (ML)	EU countries, 2005–2020	18.5%
Petrick & Wagner (2014)	Matching + DiD	German manufacturers, 2005–2010	20%
Anderson & Di Maria (2011)	Counterfactual scenarios	EU pilot phase, 2005–2007	247 $MtCO_2$ abated (pilot phase)

8. Conclusion

This study provides a rigorous evaluation of the EU ETS's estimated impact on energy sector CO_2 emissions using a staggered DiD design to address critical gaps in the existing literature. The

findings reveal that although the EU ETS has been demonstrably effective in reducing emissions, its impact varied significantly across implementation phases and participant cohorts. The 2005 cohort achieved the most substantial reductions (19.55% or -11.76 $MtCO_2$/year on average), with effects intensifying during Phase III and peaking at -24.12 $MtCO_2$/year by 2020 before moderating in subsequent years.

The analysis demonstrates that the policy's effectiveness grew as it matured, with tighter caps and auctioning mechanisms in later phases creating stronger economic incentives for decarbonization. Notably, the duration of participation emerged as a key factor, with longer exposure to the policy correlating with greater emissions reductions. Although high-emission economies like Germany and France showed particularly strong responses, later cohorts (2008 and 2013) showed smaller and often statistically insignificant aggregate effects; however, disaggregated analysis shows that the 2008 cohort's reductions were concentrated in high-emission, carbon-intensive Balkan countries, highlighting the importance of regional heterogeneity in assessing policy effectiveness. The results indicate that although the EU ETS has led to substantial progress by achieving an estimated 48% (a descriptive 1990–2022 trend reflecting post-communist deindustrialization and other structural changes, not the ETS-specific causal effect) reduction in energy sector emissions from 1990 levels by 2022, it has not yet reached the level required to meet the EU's 55% reduction target by 2030, despite being generally on track. This highlights the importance of complementary policies and measures, including national carbon taxes, ranging in Europe from $0.76 to $132.12/ton as of 2024 (World Bank Carbon Pricing Dashboard, 2025), and renewable energy incentives, to ensure sustained decarbonization beyond the minimum requirements although these fall beyond this study's scope.

Methodologically, this study makes significant contributions through its cohort-specific analysis of all EU ETS phases, employing robust econometric techniques to estimate policy effects. The 2005 cohort's results are particularly compelling, offering clear evidence of the policy's impact whereas the 2008 cohort's findings, although directionally consistent, underscore the challenges of estimating effects with smaller sample sizes. The single-country 2013 cohort (Croatia) further illustrates the limitations of analyzing small, late-adopters.

In conclusion, the EU ETS has proven to be a powerful but imperfect instrument for emissions reduction, with its effectiveness heavily dependent on policy design and implementation timing. The system's evolution, through progressively stricter caps and mechanism refinements, offers valuable lessons for carbon pricing schemes worldwide. As climate challenges intensify, these insights will be crucial for designing policies that not only reduce emissions effectively but do so equitably across all participating economies. Looking forward, further research should focus on (1)

expanding sectoral coverage beyond energy production, (2) examining policy interactions with national carbon taxes and renewable incentives, (3) incorporating post-2022 data to assess Phase IV reforms, and (4) investigating potential carbon leakage effects, considering that some countries might have moved energy production to countries with weaker regulations, causing national emissions to fall.

References

Anderson, B. and Di Maria, C. (2011). Abatement and Allocation in the Pilot Phase of the EU ETS. Environmental and Resource Economics, 48(1), 83-103.

Angrist, J.D., and Pischke, J.-S. *Mostly Harmless Econometrics*. Princeton, New Jersey: Princeton University Press, 2008.

Arkhangelsky, D., Athey, S., Hirshberg, D. A., Imbens, G. W., and Wager, S. (2021). Synthetic difference-in-differences. *American Economic Review, 111*(12), 4088–4118.

Baker, A.C., and Larcker, D. (2022). "How Much Should We Trust Staggered Difference-in-Differences Estimates?" *Journal of Financial Economics*, **144**(2): 370–395.

Bayer, P., and Aklin, M. (2020). "The European Union Emissions Trading System Reduced CO2 Emissions Despite Low Prices." *Proceedings of the National Academy of Sciences*, **117**(16): 8804–8812.

Biancalani, F., Gnecco, G., Metulini, R., and Riccaboni, M. (2024). "The Impact of the European Union Emissions Trading System on Carbon Dioxide Emissions: A Matrix Completion Analysis." *Scientific Reports*, **14**(1): 19676.

Bordignon, M., and Gamannossi degl'Innocenti, D. (2023). "Third Time's a Charm? Assessing the Impact of the Third Phase of the EU ETS on CO2 Emissions and Performance." *Sustainability*, **15**(8): 6394.

Callaway, B., and Sant'Anna, P.H. (2021). "Difference-in-Differences with Multiple Time Periods. *Journal of Econometrics*, **225**(2), 200–230.

Cameron, A.C., and Miller, D.L. (2015). "A Practitioner's Guide to Cluster-Robust Inference." *Journal of Human Resources*, **50**(2): 317–372.

Climate Watch (2025). Historical Greenhouse Gas Emissions. Washington, DC: World Resources Institute. Available online at: https://www.climatewatchdata.org/ghg-emissions?end_year=2022&start_year=1990. Dec. 2025.

Colmer, J., Martin, R., Muûls, M., and Wagner, U.J. (2025). "Does Pricing Carbon Mitigate Climate Change? Firm-Level Evidence from the European Union Emissions Trading System." *Review of Economic Studies*, **92**(3): 1625–1660.

De Chaisemartin, C., and d'Haultfoeuille, X. (2020). "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects." *American Economic Review*, **110**(9): 2964–2996.

Dechezleprêtre, A., Nachtigall, D., and Venmans, F. (2023). "The Joint Impact of the European Union Emissions Trading System on Carbon Emissions and Economic Performance." *Journal of Environmental Economics and Management*, **118**: 102758.

Demailly, D., and Quirion, P. (2008). "European Emission Trading Scheme and Competitiveness: A Case Study on the Iron and Steel Industry." *Energy Economics*, **30**(4): 2009–2027.

Ellerman, A., and Buchner, B. (2008). "Over-Allocation or Abatement? A Preliminary Analysis of the EU ETS Based on the 2005–06 Emissions Data." *Environmental and Resource Economics*, **41**: 267–287.

European Commission (2025). *EU Climate Action Progress Report 2025: Strengthening Competitiveness on the Road to Climate Neutrality*. Directorate-General for Climate Action. Luxembourg: Publications Office of the European Union.

European Environment Agency (2025). *Total GHG Emissions and Removals in the EU* (Version 3.0) [Dataset]. European Environment Agency. https://doi.org/10.2909/6331f651-8863-4656-a911-669f2a332a1e. Accessed: Dec. 2025.

Eurostat. (2025). Complete Energy Balances (nrg_bal_c) [Dataset]. Statistical Office of the European Union. https://doi.org/10.2908/nrg_bal_c [ec.europa.eu]. Accessed: Dec. 2025.

Fageda, X., and Teixidó, J.J. (2022). "Pricing Carbon in the Aviation Sector: Evidence from the European Emissions Trading System." *Journal of Environmental Economics and Management*, **111**: 102591.

Fowlie, M., Holland, S. P., and Mansur, E. T. (2012). "What Do Emissions Markets Deliver and

to Whom? Evidence from Southern California's Nox Trading Program." *American Economic Review*, **102**(2): 965–93.
Fox, J., and Weisberg, S. *An R Companion to Applied Regression*, Thousand Oaks, CA: Sage Publications, Inc., 2018.
Goodman-Bacon, A. (2021). "Difference-in-Differences with Variation in Treatment Timing." *Journal of Econometrics*, **225**(2): 254–277.
Imai, K., and Kim, I.S. (2021). "On the Use of Two-Way Fixed Effects Regression Models for Causal Inference with Panel Data." *Political Analysis*, **29**(3), 405–415.
International Energy Agency (IEA) (2025). World Energy Statistics and Balances. https://www.iea.org/reports/world-energy-balances-overview. Accessed: Dec. 2025.
James, G., Witten, D., Hastie, T., and Tibshirani, R. *An Introduction to Statistical Learning*, New York, NY: Springer, 2021.
Jaraitė, J., and Maria, C.D. (2016). "Did the EU ETS Make a Difference? An Empirical Assessment Using Lithuanian Firm-Level Data." *The Energy Journal*, **37**(2): 68–92.
Jones, M.W., Peters, G.P., Gasser, T., Andrew, R.M., Schwingshackl, C., Gütschow, J., Houghton, R.A., Friedlingstein, P., Pongratz, J., and Le Quéré, C. (2023). "National Contributions to Climate Change due to Historical Emissions of Carbon Dioxide, Methane, and Nitrous Oxide Since 1850." *Scientific Data*, **10**(1): 155.
Kara, M., Syri, S., Lehtilä, A., Helynen, S., Kekkonen, V., Ruska, M., and Forsström, J. (2008). "The Impacts of EU CO2 Emissions Trading on Electricity Markets and Electricity Consumers in Finland." *Energy Economics*, **30**(2), 193–211.
Laing, T., Sato, M., Grubb, M., Comberti, C. (2014). "The Effects and Side-Effects of the EU Emissions Trading Scheme." *Wiley Interd. Reviews: Climate Change*, **5**(4): 509–519.
Lise, W., Sijm, J., and Hobbs, B. (2010). "The Impact of the EU ETS on Prices, Profits and Emissions in the Power Sector: Simulation Results with the Competes EU20 Model." *Environmental and Resource Economics*, **47**: 23–44.
Martin, R., Muûls, M., de Preux, L.B., and Wagner, U.J. (2014). "Industry Compensation Under Relocation Risk: A Firm-Level Analysis of the EU Emissions Trading Scheme. *American Economic Review*, **104**(8): 2482–2508.
Nguyen, M. *A Guide on Data Analysis: From Basics to Causal Inference*. Cham: Springer.
Petrick, S., and Wagner, U.J. (2014). The Impact of Carbon Trading on Industry: Evidence from German Manufacturing Firms. Available at SSRN 2389800.
Roodman, D., Nielsen, M. Ø., MacKinnon, J. G., & Webb, M. D. (2019). "Fast and wild: Bootstrap inference in Stata using boottest." *The Stata Journal, 19*(1), 4–60.
Sant'Anna, P.H., and Zhao, J. (2020). "Doubly Robust Difference-in-Differences Estimators." *Journal of Econometrics*, **219**(1): 101–122.
Sun, L., & Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics, 225*(2), 175–199. https://doi.org/10.1016/j.jeconom.2020.09.006
Wing, C., Yozwiak, M., Hollingsworth, A., Freedman, S., and Simon, K. (2024). "Designing Difference-in-Difference Studies with Staggered Treatment Adoption: Key Concepts and Practical Guidelines." *Annual Review of Public Health*, **45**(1): 485–505.
World Bank (2025). Carbon Pricing Dashboard — Compliance Price [Data set]. Retrieved Dec. 2025: https://carbonpricingdashboard.worldbank.org/compliance/price.

Appendix I. Model Validation

Parallel Trends and No Anticipation

The parallel trends assumption, which requires treated and control units to follow similar trajectories before treatment in the absence of anticipation effects, is fundamental to any DiD design. In this framework, four variations of this assumption must hold: comparing treated units to never-treated and not-yet-treated countries, both unconditionally and conditional on covariates. To assess this, this study employs both visual inspection of pretreatment annual emissions and formal statistical tests. Initial observations of emissions trends before 2005 provide a preliminary indication of whether the parallel trends assumption is plausible as observed in Figure A1.

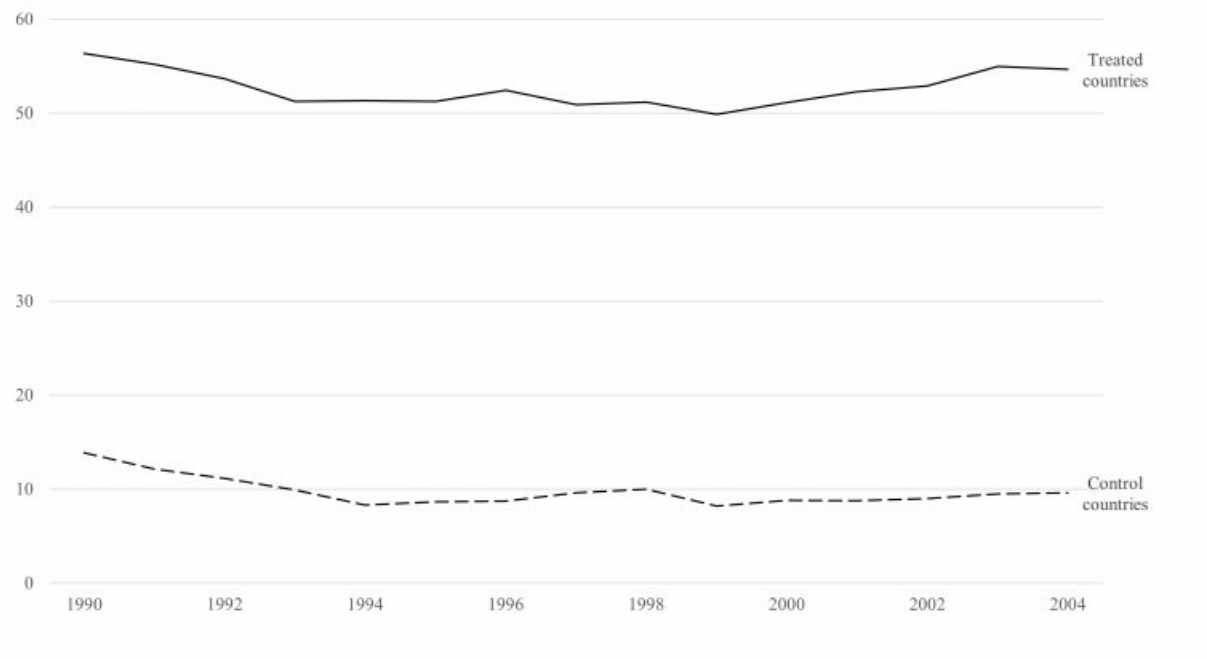

Figure A1. Average annual emissions in the energy production sector by treatment group from 1990 to 2004.

- *Model 1*

We tested the parallel trends assumption using never-treated and not-yet-treated countries as the control group, with comparisons made separately by cohorts due to the policy's staggered implementation. This specification was applied only to the 2005 and 2008 cohorts because the 2013 cohort lacked later-joining countries to serve as not-yet-treated controls. We compared pretreatment trends up to each cohort's implementation year. Visual inspection was based on plots

showing local polynomial smoothed confidence intervals of average emissions for treated (red) and control (blue) countries, providing a flexible, smooth representation of the data without assuming a specific functional form. Additionally, we presented raw averages and individual country emissions in the pretreatment period. Visually, the parallel trends assumption appears to hold across all cohorts, as can be observed in Figures A2 and A3.

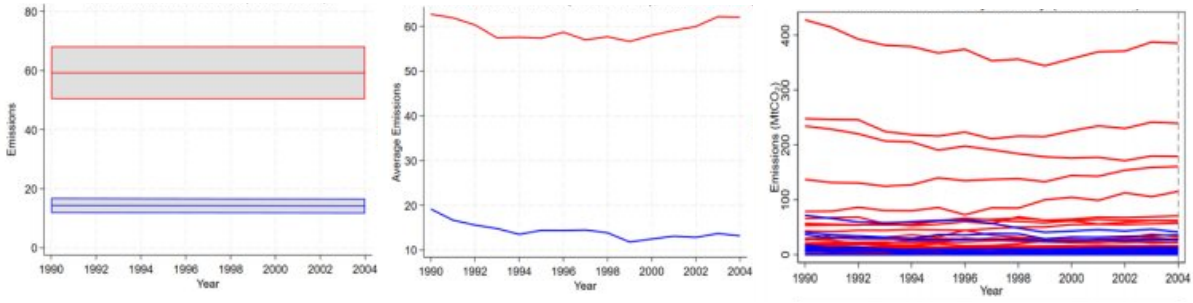

Figure A2. Visual inspection of parallel trends (1990–2004) for the 2005 cohort using model 1. Red lines represent treated countries; blue lines represent control countries. from left to right: (i) local polynomial-smoothed confidence intervals of average emissions, (ii) raw average by treatment group, and (iii) individual country emissions.

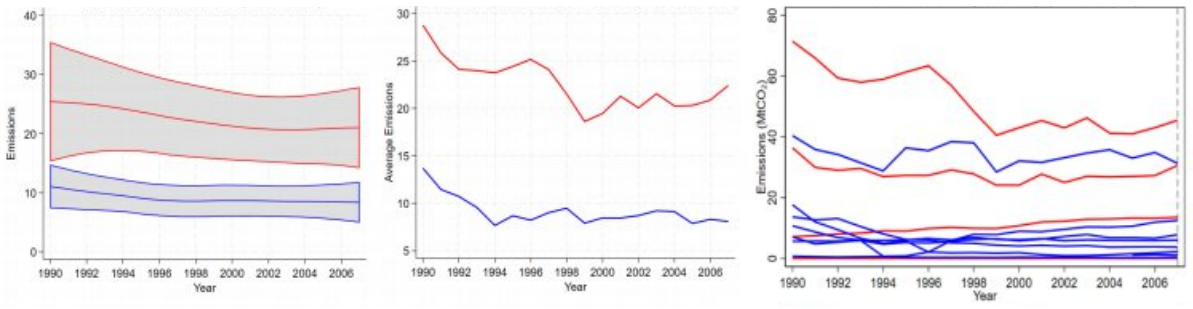

Figure A3. Visual inspection for parallel trends (1990–2007) for the 2008 cohort using model 1. Red lines represent treated countries; blue lines represent control countries. from left to right: (i) local polynomial-smoothed confidence intervals of average emissions, (ii) raw group averages by treatment status, and (iii) individual country emissions.

To statistically test for differences in pretreatment trends, we applied the approach proposed by Nguyen (2020), using only pretreatment data specific to each treatment cohort. The key coefficient of interest is the interaction term, which captures differences in time trends between treated and control groups; the statistical insignificance of this parameter suggests that the parallel trends assumption holds. The model is formally specified as

(A1) $$Yit = \alpha i + \delta t + \sum_c \sum_t \beta_{ct} \left(1_{\in Cohort_c} x\, 1_{t=Year_t}\right) + \epsilon_{it}$$

In this specification, αi represent country fixed effects, and δt represents year fixed effects.

The coefficients of interest, β_{ct}, capture the difference in time trends for the dependent variable (emissions) between treated cohorts and the control group. Year fixed effects are included as dummies rather than as a continuous time variable, allowing for year-specific treatment effects without imposing linearity. Results indicate that for the 2005 cohort, all pretreatment interaction terms were statistically insignificant at the 90% level or higher, with the exception of 2003 and 2004, which were insignificant only at the 5% level. For the 2008 cohort, all coefficients were highly insignificant, further affirming the absence of differential pretreatment trends. These findings suggest no substantial violations of the parallel-trends assumption when using never-treated and not-yet-treated countries as the control group. Together with visual evidence, the parallel trends assumption is satisfied in Model 1.

- *Model 3*

We conducted a similar analysis using only never-treated countries as the control group. This specification was applied to all cohorts.. Visually, the smoothed and raw-average emissions plots (Figures A4 and A5) remained largely unchanged although we observed a slightly less smoother trend for the 2008 cohort. Nonetheless, the graphs continued to support the parallel trends assumption. The statistical test following Nguyen (2020) yielded consistent results with the previous specification. Results indicate that for the 2005 and 2008 cohorts, all interaction terms were statistically insignificant at the 90% confidence level or higher. For the 2013 cohort, most coefficients were insignificant at the 10% level, with some marginally significant at the 5%–10% level. These findings, along with the visual inspection, reinforce the validity of the DiD design under this expanded control group definition.

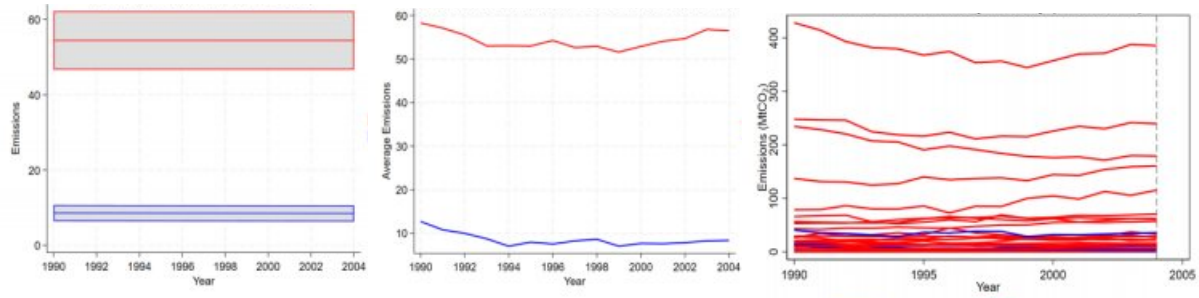

Figure A4. Visual inspection for parallel trends (1990–2004) for the 2005 cohort using model 3. Red lines represent treated countries; blue lines represent control countries. from left to right: (i) local polynomial-smoothed confidence intervals of average emissions, (ii) raw average by treatment group, and (iii) individual country emissions.

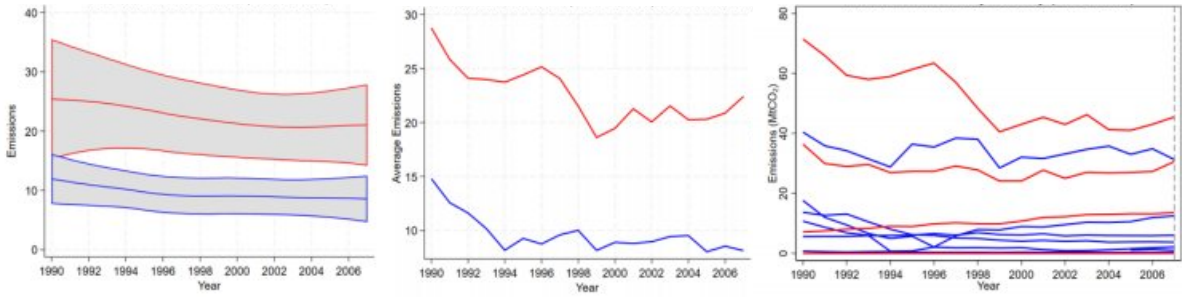

Figure A5. Visual inspection for parallel trends (1990–2007) for the 2008 cohort using model 3 Red lines represent treated countries; blue lines represent control countries. from left to right: (i) local polynomial-smoothed confidence intervals of average emissions, (ii) raw average by treatment group, and (iii) individual country emissions.

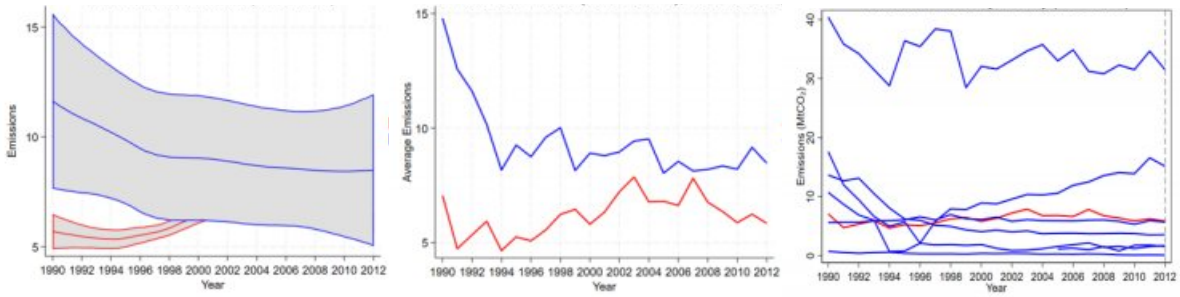

Figure A6. Visual inspection for parallel trends (1990–2012) for the 2013 cohort using model 3. Red lines represent treated countries; blue lines represent control countries. from left to right: (i) local polynomial-smoothed confidence intervals of average emissions, (ii) raw average by treatment group, and (iii) individual country emissions.

- *Model 2*

We further tested the parallel trends assumption using never-treated and not-yet-treated countries as the control group, conditional on the renewable energy share. As before, analysis was restricted to the 2005 and 2008 cohorts. To implement this, emissions were first residualized using a fixed-effects regression that controls for the share of renewable energy. This approach isolates the variation in emissions that is orthogonal to the covariate, thereby removing the confounding influence of renewable energy adoption. Figure A7 presents the average residualized emission trends for treated and control groups. This procedure allows for an evaluation of whether parallel trends hold in the residual variation, representing a weaker but still informative version of the parallel trends assumption.

To facilitate interpretation, we added the average of the original emissions back to the residuals, restoring the values to the $MtCO_2$ scale rather than deviations from zero. Subsequently, for each treatment cohort (2005 and2008), the local polynomial smoothed confidence intervals, raw averages,

and cohort-specific average emissions were plotted in Figures A8 and A 9 .

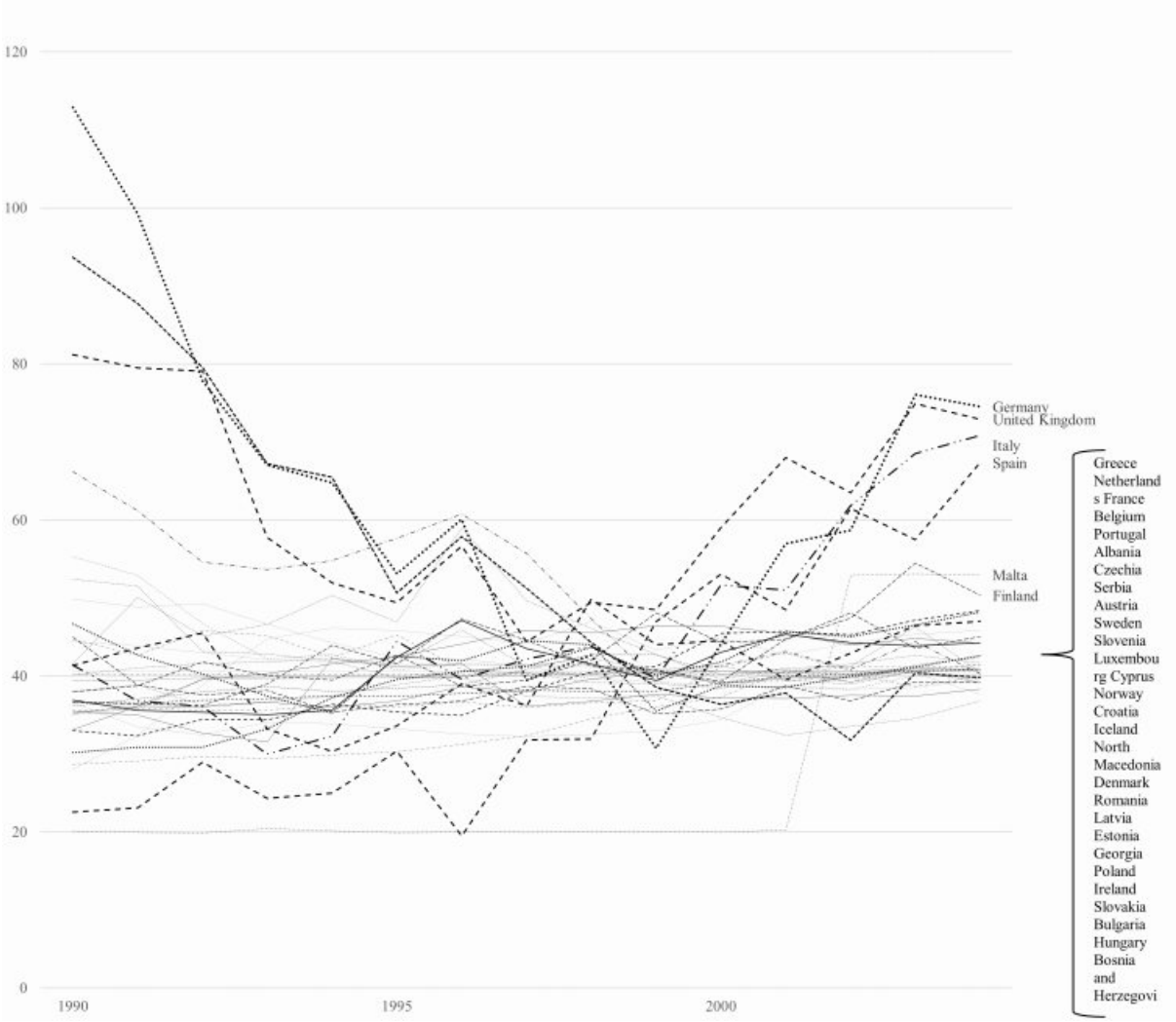

Figure A7. Residualized annual emissions in the energy production sector by country from 1990 to 2004.

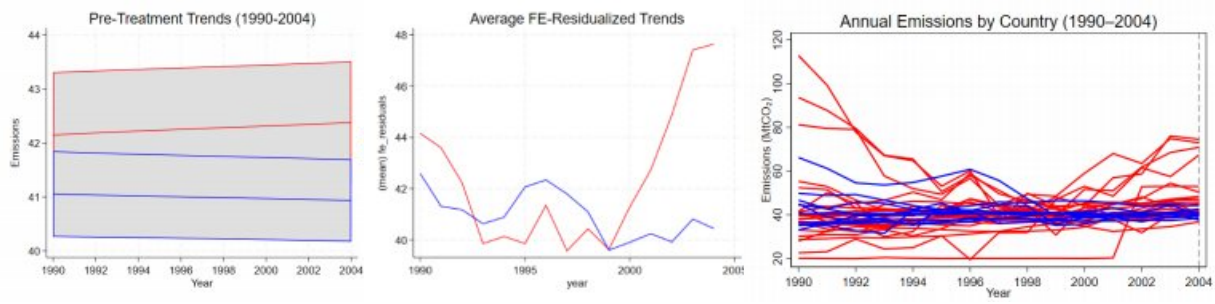

Figure A8. Visual inspection for parallel trends (1990–2004) for the 2005 cohort using model 2 Red lines represent treated countries; blue lines represent control countries. from left to right: (i) local polynomial–smoothed confidence intervals of average emissions, (ii) raw average by treatment group, and (iii) individual country emissions.

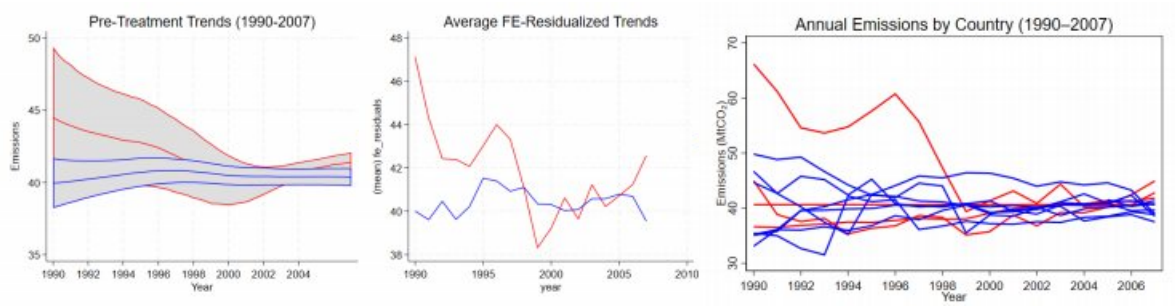

Figure A9. visual inspection for parallel trends (1990–2007) for the 2008 cohort using model 2. Red lines represent treated countries; blue lines represent control countries. from left to right: (i) local polynomial-smoothed confidence intervals of average emissions, (ii) raw average by treatment group, and (iii) individual country emissions.

When controlling for the renewable energy share, visual inspection of residualized emission trends suggested potential violations of the parallel trends assumption. However, formal statistical analysis based on Nguyen's (2020) pretreatment trend-differences approach offers a more nuanced perspective. Using a modified regression model:

(A2) $$\tilde{Y}it = \alpha i + \delta t + \sum_c \sum_t \beta_{ct} \ (1_{\in Cohort_c} x\ 1_{t=Year_t} x R_{it}\} + \epsilon_{it}$$

In this specification, $\tilde{Y}_{it}$denotes the residualized emissions, controlling for covariates other than the renewable energy share; $\alpha_i$represents country fixed effects; and $\delta_t$represents year fixed effects. R_{it}corresponds to the renewable energy share in country i at time t. The coefficients of interest, β_{ct}, capture the differential pretreatment trends conditional on the renewable energy share.. For the 2005 cohort, all triple interaction coefficients remained statistically insignificant at the 90% confidence level or higher, indicating no violation of the parallel trends assumption when using this expanded control group. However, for the 2008 cohort, several triple interaction terms were statistically significant at the 95% level, suggesting a breach of the parallel trends assumption and raising concerns about potential bias in the estimated treatment effects for this group. Notably, in the main model estimations, treatment effects for the 2008 cohort were consistently insignificant across all specifications, including those that satisfied all identifying assumptions. As such, the detected violation did not materially affect the robustness of the study's overall conclusions.

- *Model 4*

Finally, the same approach as in Model 2 is extended by only introducing never treated countries into the control group. This adjustment yielded results broadly consistent with the previous specification.

Although minor concerns persisted regarding the visual plausibility of the parallel trends assumption, particularly in the 2013 cohort. In the 2005 and 2008 cohorts, all interaction terms were statistically insignificant at the 90% confidence level or higher. For the 2013 cohort, most coefficients remained insignificant above the 10% level, with only a few marginal cases falling between 5%–10%.

This contrast between visual and statistical findings reflects differing sensitivities: whereas visual plots may suggest divergence due to random noise or small sample sizes, statistical tests more rigorously evaluate whether differences are significant. Thus, despite slight visual deviations, the formal analysis does not reject the parallel trends assumption, reinforcing the robustness of the identification strategy.

Overlap

Overlap, a key assumption in the Callaway and Sant'Anna (2021) framework, requires that for each treatment cohort *g*, the probability of receiving treatment in period *t* lies strictly between 0 and 1. This ensures that treated and control units are comparable, particularly in specifications involving covariates, in which the doubly robust method relies on balancing propensity scores to adjust for differences. In models without covariates, this assumption is less critical because ATT identification relies solely on group-level variation. To test overlap in the covariate-adjusted models, propensity scores were estimated via logistic regression. For the 2005 cohort, using both never-treated and not-yet-treated countries as controls, treated and control groups exhibited nearly identical scores with minimal variation, indicating strong comparability. In the 2008 cohort, mean propensity scores differed moderately (0.425 vs. 0.373) but exhibited sufficient variation and acceptable overlap, offering both predictive strength and balance, making it the most robust cohort. By contrast, the 2013 cohort included only one treated country, resulting in severe limitations: minimal variation, poor model fit, and a lack of common support. Although the 2013 estimates were not considered reliable, they still contributed informative context for the overall analysis. Results are plotted in Figure A11.

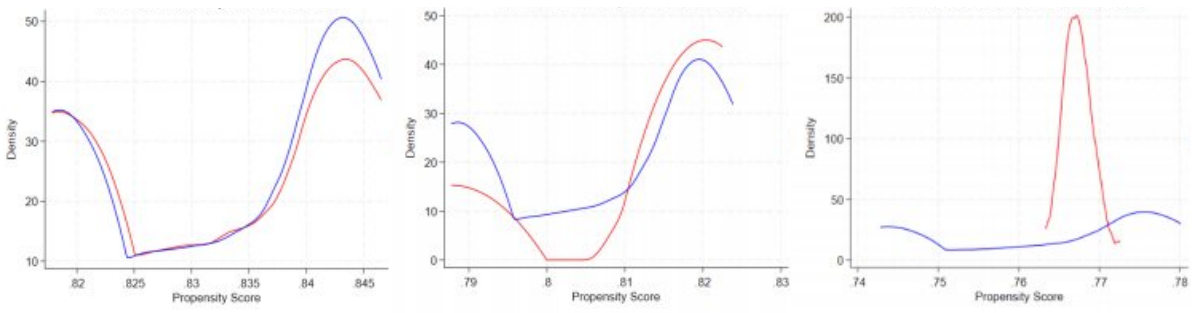

Figure A11. Visual inspection for overlap using propensity scores distribution in model 2 for 2005, 2008, and 2013 cohorts. Red lines represent treated countries; blue lines represent eligible controls.

When testing using as the control group never-treated countries, the results remained largely consistent with the previous specification. For the 2005 and 2008 cohorts, the propensity scores for treated and control groups remained nearly identical, indicating well-balanced comparisons across groups. Notably, the 2008 cohort continued to exhibit the strongest predictive power and overlap. Figure A11 and Figure A12 together show that incorporating not-yet-treated countries alongside never-treated ones enhanced the robustness of the specification, resulting in a notably stronger and more reliable model compared with the latter version.

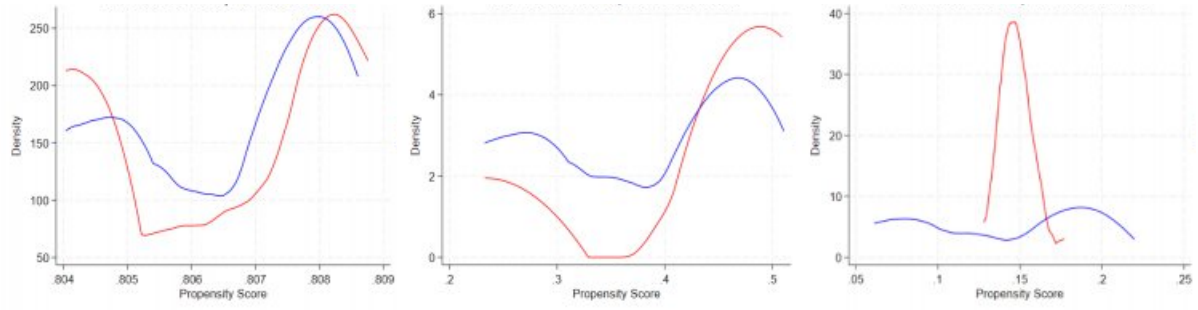

Figure A12. Visual inspection for overlap using propensity scores distribution in model 4for 2005, 2008, and 2013 cohorts. Red lines represent treated countries; blue lines represent eligible controls.

Appendix II. Additional robustness checks

Log(CO_2) Specification

We re-estimate all specifications using the natural logarithm of CO_2 emissions as the outcome variable. This transformation allows effects to be interpreted as approximate percentage changes and reduces sensitivity to very large emitters.

The results closely align with the baseline estimates. The overall ATT ranges from -0.218 to -0.238 log points, corresponding to an approximate 19.6%–21.2% reduction, which is consistent with the 17.5%–19.6% effect obtained in the levels specification. For the 2005 cohort, effects range from -0.197 to -0.217 log points and are weakly significant in Models 1 and 2, implying reductions of roughly 18%–20%, with effects increasing over time to about -0.49 to -0.52 log points by 2019–2020 ($P < 0.01$), consistent with the compounding pattern found in the main analysis.

For the 2008 cohort, average effects remain statistically insignificant across all specifications. In contrast, the 2013 cohort shows a stable and significant reduction of approximately -0.205 to -0.207 log points ($P < 0.01$), corresponding to a 19% decrease. The subgroup analysis for the 2008 cohort confirms heterogeneous effects: while Nordic countries show negative but insignificant effects, Balkan countries exhibit statistically significant reductions from 2019 onward, reaching -0.402 to -0.538 log points, equivalent to roughly 33%–42% decreases.

Table A1—Re-estimation of estimates using Log(CO_2).

	Model (1) $MtCO_2$	Model (2) $MtCO_2$	Model (3) $MtCO_2$	Model (4) $MtCO_2$
g2005				
t_2004_2005	-0.0522	-0.0702*	-0.0703	-0.0854*
	(0.0319)	(0.0374)	(0.0486)	(0.0472)
t_2004_2006	-0.158	-0.240*	-0.0967	-0.113
	(0.0962)	(0.122)	(0.0731)	(0.0785)
t_2004_2007	-0.327*	-0.474**	-0.165*	-0.191*
	(0.177)	(0.237)	(0.0971)	(0.102)
t_2004_2008	-0.121**	-0.127***	-0.130**	-0.134**
	(0.0525)	(0.0475)	(0.0594)	(0.0542)
t_2004_2009	-0.0762	-0.0785	-0.0789	-0.0804
	(0.127)	(0.137)	(0.147)	(0.154)
t_2004_2010	0.0662	0.0763	0.0681	0.0766
	(0.148)	(0.157)	(0.172)	(0.178)
t_2004_2011	-0.104	-0.0916	-0.112	-0.100
	(0.146)	(0.155)	(0.170)	(0.175)
t_2004_2012	-0.0778	-0.0786	-0.0878	-0.0873
	(0.167)	(0.182)	(0.194)	(0.204)
t_2004_2013	0.0980	0.134	0.0980	0.134
	(0.269)	(0.271)	(0.269)	(0.271)
t_2004_2014	-0.103	-0.101	-0.103	-0.101
	(0.140)	(0.140)	(0.140)	(0.140)
t_2004_2015	-0.128	-0.128	-0.128	-0.128
	(0.206)	(0.214)	(0.206)	(0.214)
t_2004_2016	-0.219	-0.215	-0.219	-0.215
	(0.175)	(0.169)	(0.175)	(0.169)
t_2004_2017	-0.339***	-0.328***	-0.339***	-0.328***
	(0.122)	(0.109)	(0.122)	(0.109)

t_2004_2018	-0.265*	-0.256*	-0.265*	-0.256*
	(0.158)	(0.149)	(0.158)	(0.149)
t_2004_2019	-0.488***	-0.492***	-0.488***	-0.492***
	(0.123)	(0.120)	(0.123)	(0.120)
t_2004_2020	-0.516***	-0.521***	-0.516***	-0.521***
	(0.169)	(0.163)	(0.169)	(0.163)
t_2004_2021	-0.443***	-0.450***	-0.443***	-0.450***
	(0.153)	(0.142)	(0.153)	(0.142)
t_2004_2022	-0.505***	-0.505***	-0.505***	-0.505***
	(0.172)	(0.178)	(0.172)	(0.178)
g2008				
t_2007_2008	-0.192	-0.216	-0.209	-0.239
	(0.212)	(0.169)	(0.216)	(0.169)
t_2007_2009	-0.148	-0.174	-0.150	-0.183
	(0.269)	(0.211)	(0.276)	(0.215)
t_2007_2010	-0.137	-0.167	-0.151	-0.189
	(0.290)	(0.235)	(0.305)	(0.248)
t_2007_2011	-0.249	-0.283	-0.269	-0.313
	(0.340)	(0.275)	(0.350)	(0.281)
t_2007_2012	-0.205	-0.223	-0.226	-0.251
	(0.294)	(0.257)	(0.306)	(0.265)
t_2007_2013	-0.222	-0.301	-0.222	-0.301
	(0.442)	(0.332)	(0.442)	(0.332)
t_2007_2014	-0.252	-0.284	-0.252	-0.284
	(0.364)	(0.312)	(0.364)	(0.312)
t_2007_2015	-0.283	-0.304	-0.283	-0.304
	(0.429)	(0.390)	(0.429)	(0.390)
t_2007_2016	-0.499	-0.538	-0.499	-0.538
	(0.508)	(0.443)	(0.508)	(0.443)
t_2007_2017	-0.582	-0.639	-0.582	-0.639
	(0.504)	(0.415)	(0.504)	(0.415)
t_2007_2018	-0.556	-0.603	-0.556	-0.603
	(0.501)	(0.428)	(0.501)	(0.428)
t_2007_2019	-0.520	-0.547*	-0.520	-0.547*
	(0.339)	(0.301)	(0.339)	(0.301)
t_2007_2020	-0.727	-0.752*	-0.727	-0.752*
	(0.456)	(0.419)	(0.456)	(0.419)

t_2007_2021	-0.592	-0.618	-0.592	-0.618
	(0.419)	(0.379)	(0.419)	(0.379)
t_2007_2022	-0.305	-0.334	-0.305	-0.334
	(0.254)	(0.233)	(0.254)	(0.233)
g2013				
t_2012_2013	0.122	0.127	0.122	0.127
	(0.0998)	(0.0887)	(0.0998)	(0.0887)
t_2012_2014	-0.0740	-0.0726	-0.0740	-0.0726
	(0.0639)	(0.0636)	(0.0639)	(0.0636)
t_2012_2015	-0.0752	-0.0760	-0.0752	-0.0760
	(0.0546)	(0.0541)	(0.0546)	(0.0541)
t_2012_2016	-0.0606	-0.0590	-0.0606	-0.0590
	(0.0867)	(0.0856)	(0.0867)	(0.0856)
t_2012_2017	-0.246*	-0.240*	-0.246*	-0.240*
	(0.144)	(0.136)	(0.144)	(0.136)
t_2012_2018	-0.300***	-0.297***	-0.300***	-0.297***
	(0.0994)	(0.0951)	(0.0994)	(0.0951)
t_2012_2019	-0.409***	-0.408***	-0.409***	-0.408***
	(0.106)	(0.103)	(0.106)	(0.103)
t_2012_2020	-0.400***	-0.399***	-0.400***	-0.399***
	(0.0854)	(0.0843)	(0.0854)	(0.0843)
t_2012_2021	-0.336***	-0.335***	-0.336***	-0.335***
	(0.118)	(0.118)	(0.118)	(0.118)
t_2012_2022	-0.294***	-0.293***	-0.294***	-0.293***
	(0.0488)	(0.0471)	(0.0488)	(0.0471)
Observations	1203	1203	1203	1203

Note: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Standard errors in parentheses. Sample excludes years 2008, 2009, and 2010. Estimates calculated using Callaway & Sant'Anna (2021).

Interaction-Weighted Event Study (Sun & Abraham, 2021)

We re-estimate the event study using the interaction-weighted (IW) estimator proposed by Sun and Abraham (2021), which addresses the same TWFE bias problem as the Callaway and Sant'Anna (2021) approach but through a different identification strategy. Rather than defining group-time ATTs and aggregating them explicitly, Sun and Abraham (2021) interact cohort dummies with relative-time dummies and recover heterogeneity-robust estimates using cohort-share weights. This provides an independent check that the main results do not depend on the specific assumptions embedded in the preferred estimator.

Figure A13 reports the interaction-weighted event study for the levels specification, using

never-treated countries as the control group. The pre-treatment period shows coefficients that are close to zero and statistically insignificant from t = −2 onward, consistent with the parallel trends evidence reported in Appendix I. The wider confidence intervals in the earlier pre-treatment periods reflect greater estimation uncertainty at longer horizons.

In the post-treatment period, the IW estimates display a pattern that closely mirrors the dynamic estimates from the preferred specification. Effects are small and insignificant in the first two years after treatment onset, then grow steadily and become statistically significant from t = +3 onward, reaching −17.61 $MtCO_2$ at t = +13 (P < 0.01) and −25.26 $MtCO_2$ at t = +15 (P < 0.01). The simple post-treatment average across all 18 post-treatment periods is −11.35 $MtCO_2$, compared with the aggregate ATT of −10.75 $MtCO_2$ reported in Section 5.1, a difference of less than 0.6 $MtCO_2$, well within the margin of sampling uncertainty. The trajectory, the magnitude, and the pattern of increasing effects over time are all consistent across both estimators.

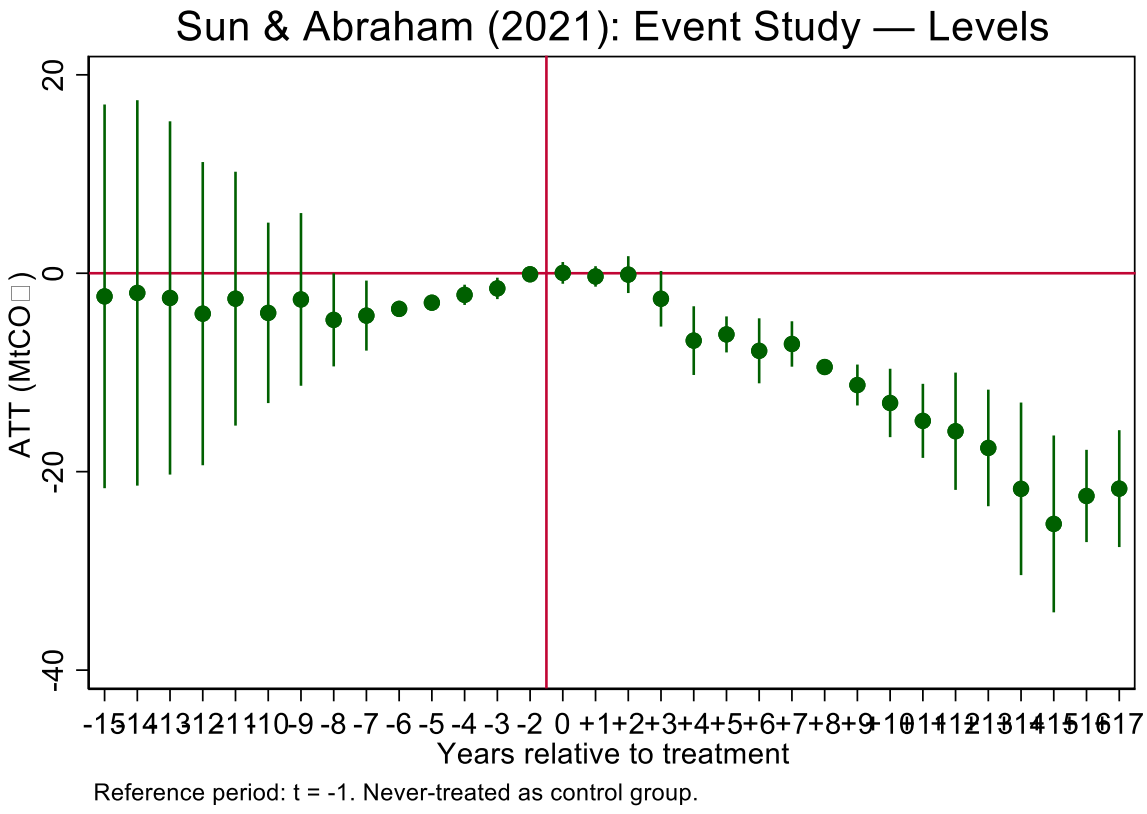

Figure A13. Sun & Abraham (2021) estimator results from Model 1.

Wild Cluster Bootstrap Inference

As noted in the discussion of clustering choices above, our sample includes 37 cross-sectional units, a relatively limited number of clusters for which asymptotic cluster-robust standard errors may not provide reliable inference (Cameron and Miller, 2015). To address this concern formally,

we implement the wild cluster bootstrap (Roodman et al., 2019) using the boottest package in Stata, applying 999 replications with Rademacher weights. Because the doubly robust staggered DiD estimator does not admit direct reestimation under bootstrap resampling, we instead apply the wild cluster bootstrap to a linear two-way fixed effects (TWFE) specification restricted to each cohort and the never-treated control group. Restricting the sample in this way ensures that the TWFE coefficient is identified solely from treated-versus-never-treated comparisons.

This restricted specification is not numerically equivalent to the corresponding Callaway and Sant'Anna group-time estimate, since the latter aggregates period-specific effects using cohort-size weighting and, in some specifications, doubly robust propensity-score adjustment; nonetheless, it offers a methodologically cleaner and more directly informative approximation of the cluster-level uncertainty surrounding each cohort comparison than the unrestricted TWFE would provide. Table A2 reports conventional cluster-robust p-values alongside their wild cluster bootstrap counterparts.

The results are reassuring. For the overall sample without covariates, the conventional p-value of 0.0653 is close to the bootstrap p-value of 0.0711, both marginally significant at the 10% level. With covariates, both the conventional (0.3667) and bootstrap (0.3624) p-values indicate an insignificant effect, in agreement. For the 2005 cohort, the effect remains significant under both approaches, although the bootstrap p-value (0.0180) is somewhat larger than the conventional one (0.0099), suggesting that asymptotic inference modestly overstates the precision of this estimate but does not overturn its significance at conventional levels.

For the 2008 cohort and its Nordic and Balkan subgroups, both conventional and bootstrap p-values indicate statistically insignificant effects, consistent with the main results using the TWFE. Notably, the bootstrap p-values for these comparisons are systematically larger than their conventional counterparts (2008 cohort: 0.4704 vs. 0.6877; Nordic: 0.2755 vs. 0.3614; Balkan: 0.2339 vs. 0.4945), reflecting the additional uncertainty introduced by the small number of treated units in these comparisons (four countries for the full 2008 cohort, two for each subgroup). For the 2013 cohort the bootstrap p-value (0.4688) is likewise larger than the conventional one (0.3118), consistent with the limitations of single-unit cohorts already discussed in Section 5.3 and Appendix I.

Overall, the wild cluster bootstrap confirms the pattern of statistical significance reported throughout the paper: the 2005 cohort and the overall ATT remain significant, while the 2008 cohort, its subgroups, and the 2013 cohort remain statistically insignificant or only marginally so.

Table A2—Wild Cluster Bootstrap Inference: Conventional vs. Bootstrap P-values

Specification	Conventional p-value	Wild Cluster Bootstrap p-value	Significance (conv.)	Significance (boot.)
Overall (no covariates)	0.0653	0.0711	*	*
Overall (with covariates)	0.3667	0.3624		
Cohort 2005	0.0099	0.0180	***	**
Cohort 2008	0.4704	0.6877		
Cohort 2013	0.3118	0.4688		
2008 Nordic subgroup	0.2755	0.3614		
2008 Balkan subgroup	0.2339	0.4945		

Leave-One-Out Sensitivity Analysis

We conduct a leave-one-out sensitivity analysis. This diagnostic exercise sequentially drops one control country at a time from the never-treated pool (Albania, Bosnia and Herzegovina, Georgia, Moldova, Montenegro, North Macedonia, and Serbia) and re-estimates the aggregate Average Treatment Effect on the Treated (ATT).

As shown in Figure A14, the estimated ATT shows stability across all iterations, remaining tightly bounded around the baseline estimate of -10.76 MtC02 (represented by the red diamond). The point estimates range from a minimum point effect of -10.11 MtCO2 when excluding Bosnia and Herzegovina to a maximum point effect of -11.25 MtCO2 when excluding Serbia. Crucially, across all iterations, the point estimates remain statistically significant at the 1% level, and their associated 95% confidence intervals overlap completely with the baseline model. This confirms that our main findings reflect a systemic policy effect across the treated cohorts rather than idiosyncratic emissions trends or data anomalies within a single control jurisdiction.

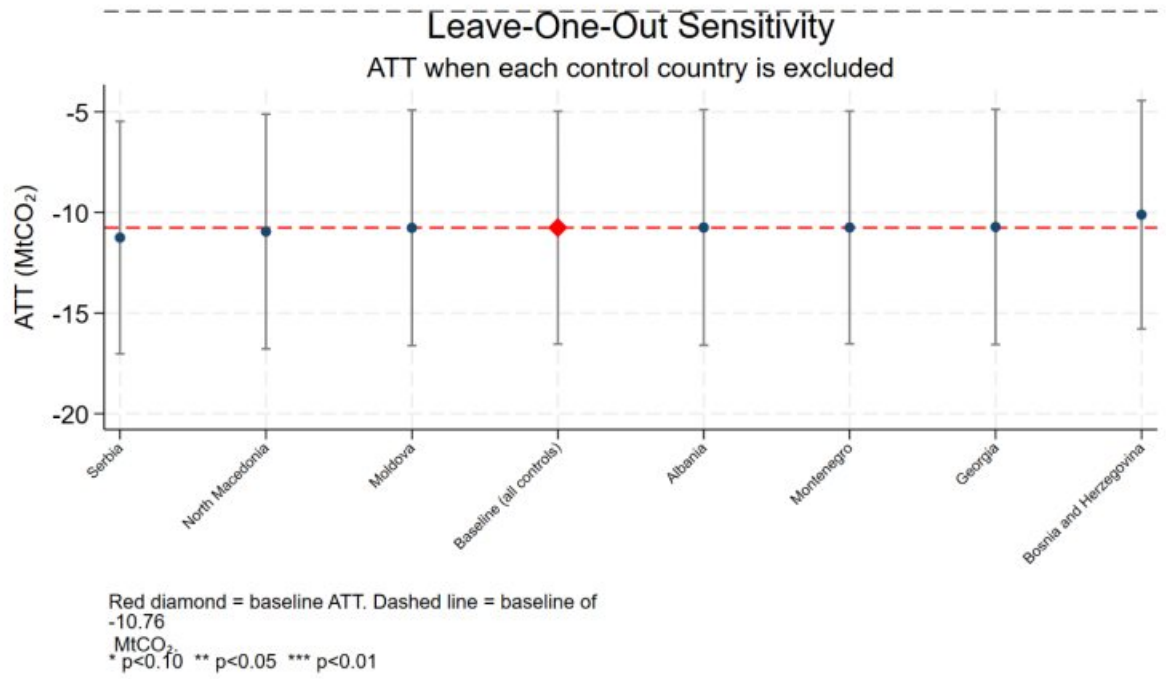

Figure A14: *Leave-One-Out Sensitivity Analysis using Model 1*

Covid-19 Period Exclusion

We re-estimate the models on a sample restricted to 1990–2019, excluding the COVID-19 period. The purpose is to assess whether the main results are driven by pandemic-related economic disruption.

The estimates closely replicate the baseline findings. The 2005 cohort shows a steadily increasing and statistically significant negative effect, reaching similar magnitudes by 2019 as in the full sample. The 2008 cohort remains statistically insignificant throughout, while the 2013 cohort retains a stable and significant negative effect with virtually unchanged magnitude. Pretreatment coefficients remain mostly insignificant, supporting the identifying assumptions.

Overall, the results indicate that the main effects are not driven by the COVID-19 period, as the observed trajectory is already fully established by 2019.

Table A3—Robustness: Sample Restricted to 2019 (Excluding COVID-19 Period)

	Model (1) $MtCO_2$	Model (2) $MtCO_2$	Model (3) $MtCO_2$	Model (4) $MtCO_2$
g2005				
t_2004_2005	-0.121	-0.205	0.0533	-0.0521
	(0.792)	(0.769)	(0.872)	(0.816)

t_2004_2006	-0.194	-0.154	-0.0909	-0.107
	(0.663)	(0.645)	(0.699)	(0.679)
t_2004_2007	-0.727	-0.607	0.241	0.114
	(0.871)	(0.733)	(1.016)	(0.890)
t_2004_2008	-2.829**	-2.919***	-2.800**	-2.878**
	(1.183)	(1.020)	(1.287)	(1.125)
t_2004_2009	-7.700***	-7.740***	-7.772***	-7.800***
	(2.420)	(2.355)	(2.460)	(2.395)
t_2004_2010	-6.334***	-6.411***	-6.430***	-6.487***
	(2.421)	(2.337)	(2.466)	(2.382)
t_2004_2011	-8.662***	-8.504***	-8.863***	-8.716***
	(2.531)	(2.520)	(2.577)	(2.567)
t_2004_2012	-7.690***	-7.728***	-7.837***	-7.858***
	(2.112)	(2.007)	(2.180)	(2.078)
t_2004_2013	-10.34***	-10.27***	-10.34***	-10.27***
	(3.004)	(2.979)	(3.004)	(2.979)
t_2004_2014	-12.41***	-12.67***	-12.41***	-12.67***
	(3.987)	(3.755)	(3.987)	(3.755)
t_2004_2015	-14.07***	-14.14***	-14.07***	-14.14***
	(4.277)	(4.222)	(4.277)	(4.222)
t_2004_2016	-16.07***	-16.07***	-16.07***	-16.07***
	(5.217)	(5.184)	(5.217)	(5.184)
t_2004_2017	-16.92***	-16.86***	-16.92***	-16.86***
	(5.753)	(5.759)	(5.753)	(5.759)
t_2004_2018	-19.00***	-19.02***	-19.00***	-19.02***
	(6.301)	(6.263)	(6.301)	(6.263)
t_2004_2019	-23.91***	-23.96***	-23.91***	-23.96***
	(7.643)	(7.599)	(7.643)	(7.599)
g2008				
t_2007_2008	0.0940	0.0374	-0.0441	-0.132
	(0.570)	(0.589)	(0.564)	(0.585)
t_2007_2009	-2.208	-2.329	-2.416	-2.595
	(1.886)	(1.973)	(1.881)	(1.978)
t_2007_2010	-2.447	-2.518	-2.699	-2.820
	(2.807)	(2.846)	(2.801)	(2.847)
t_2007_2011	-1.128	-1.476	-1.452	-1.923
	(2.452)	(2.573)	(2.455)	(2.582)
t_2007_2012	-2.489	-2.612	-2.777	-2.968
	(2.446)	(2.531)	(2.438)	(2.538)
t_2007_2013	-5.301	-5.649	-5.301	-5.649
	(3.599)	(3.824)	(3.599)	(3.824)
t_2007_2014	-3.461	-3.257	-3.461	-3.257
	(3.739)	(3.692)	(3.739)	(3.692)

t_2007_2015	-4.077	-4.165	-4.077	-4.165
	(3.845)	(3.893)	(3.845)	(3.893)
t_2007_2016	-5.861	-6.083	-5.861	-6.083
	(4.224)	(4.374)	(4.224)	(4.374)
t_2007_2017	-5.407	-5.745	-5.407	-5.745
	(4.128)	(4.328)	(4.128)	(4.328)
t_2007_2018	-6.584	-6.791	-6.584	-6.791
	(4.388)	(4.534)	(4.388)	(4.534)
t_2007_2019	-7.656	-7.781	-7.656	-7.781
	(5.014)	(5.093)	(5.014)	(5.093)
g2013				
t_2012_2013	-0.493*	-0.473**	-0.493*	-0.473**
	(0.295)	(0.239)	(0.295)	(0.239)
t_2012_2014	0.238	0.197	0.238	0.197
	(0.633)	(0.516)	(0.633)	(0.516)
t_2012_2015	-0.725	-0.737*	-0.725	-0.737*
	(0.444)	(0.415)	(0.444)	(0.415)
t_2012_2016	-0.798**	-0.795**	-0.798**	-0.795**
	(0.395)	(0.380)	(0.395)	(0.380)
t_2012_2017	-1.336***	-1.319***	-1.336***	-1.319***
	(0.418)	(0.378)	(0.418)	(0.378)
t_2012_2018	-1.614***	-1.611***	-1.614***	-1.611***
	(0.311)	(0.300)	(0.311)	(0.300)
t_2012_2019	-1.612***	-1.619***	-1.612***	-1.619***
	(0.287)	(0.271)	(0.287)	(0.271)
Observations	1095	1095	1095	1095

Note:* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses. Sample excludes years 2008, 2009, and 2010. Estimates calculated using Callaway & Sant'Anna (2021).

Excluding the Global Financial Crisis (2008–2010)

We address the potential confounding effect of the Global Financial Crisis (2008–2010) by excluding these years from the sample and re-estimating the model. This checks whether the observed emissions reductions are driven by recessionary demand shocks rather than EU ETS policy.

The results, reported in Table A3 remain highly consistent with the baseline estimates. For the 2005 cohort, significant reductions emerge immediately after treatment and grow over time, reaching approximately -23.9 $MtCO_2$ by 2019 ($p < 0.01$), closely mirroring the main specification. The 2013 cohort similarly exhibits stable and statistically significant reductions throughout the post-treatment period, while the 2008 cohort shows persistently negative but less precisely estimated effects.

Importantly, pre-treatment coefficients remain flat and statistically insignificant, indicating no evidence of differential pre-trends once the crisis years are excluded.

Table A4—Robustness Check: Excluding the Global Financial Crisis (2008-2010)

	Model (1) $MtCO_2$	Model (2) $MtCO_2$	Model (3) $MtCO_2$	Model (4) $MtCO_2$
g2005				
t_2004_2011	-8.863***	-8.662***	-8.716***	-8.504***
	(2.577)	(2.531)	(2.567)	(2.520)
t_2004_2012	-7.837***	-7.690***	-7.858***	-7.728***
	(2.180)	(2.112)	(2.078)	(2.007)
t_2004_2013	-10.34***	-10.34***	-10.27***	-10.27***
	(3.004)	(3.004)	(2.979)	(2.979)
t_2004_2014	-12.41***	-12.41***	-12.67***	-12.67***
	(3.987)	(3.987)	(3.755)	(3.755)
t_2004_2015	-14.07***	-14.07***	-14.14***	-14.14***
	(4.277)	(4.277)	(4.222)	(4.222)
t_2004_2016	-16.07***	-16.07***	-16.07***	-16.07***
	(5.217)	(5.217)	(5.184)	(5.184)
t_2004_2017	-16.92***	-16.92***	-16.86***	-16.86***
	(5.753)	(5.753)	(5.759)	(5.759)
t_2004_2018	-19.00***	-19.00***	-19.02***	-19.02***
	(6.301)	(6.301)	(6.263)	(6.263)
t_2004_2019	-23.91***	-23.91***	-23.96***	-23.96***
	(7.643)	(7.643)	(7.599)	(7.599)
t_2004_2020	-24.12***	-24.12***	-24.16***	-24.16***
	(7.777)	(7.777)	(7.742)	(7.742)
t_2004_2021	-21.27***	-21.27***	-21.43***	-21.43***
	(6.797)	(6.797)	(6.713)	(6.713)
t_2004_2022	-20.54***	-20.54***	-20.63***	-20.63***
	(6.180)	(6.180)	(6.112)	(6.112)
g2008				
t_2007_2011	-1.452	-1.128	-1.923	-1.476
	(2.455)	(2.452)	(2.582)	(2.573)
t_2007_2012	-2.777	-2.489	-2.968	-2.612
	(2.438)	(2.446)	(2.538)	(2.531)
t_2007_2013	-5.301	-5.301	-5.649	-5.649
	(3.599)	(3.599)	(3.824)	(3.824)

t_2007_2014	-3.461	-3.461	-3.257	-3.257
	(3.739)	(3.739)	(3.692)	(3.692)
t_2007_2015	-4.077	-4.077	-4.165	-4.165
	(3.845)	(3.845)	(3.893)	(3.893)
t_2007_2016	-5.861	-5.861	-6.083	-6.083
	(4.224)	(4.224)	(4.374)	(4.374)
t_2007_2017	-5.407	-5.407	-5.745	-5.745
	(4.128)	(4.128)	(4.328)	(4.328)
t_2007_2018	-6.584	-6.584	-6.791	-6.791
	(4.388)	(4.388)	(4.534)	(4.534)
t_2007_2019	-7.656	-7.656	-7.781	-7.781
	(5.014)	(5.014)	(5.093)	(5.093)
t_2007_2020	-10.03*	-10.03*	-10.22*	-10.22*
	(5.728)	(5.728)	(5.873)	(5.873)
t_2007_2021	-8.730	-8.730	-8.764	-8.764
	(5.483)	(5.483)	(5.517)	(5.517)
t_2007_2022	-7.888	-7.888	-7.994	-7.994
	(5.657)	(5.657)	(5.723)	(5.723)
g2013				
t_2011_2012	0.290	0.290	0.369	0.369
	(0.420)	(0.420)	(0.359)	(0.359)
t_2012_2013	-0.493*	-0.493*	-0.473**	-0.473**
	(0.295)	(0.295)	(0.239)	(0.239)
t_2012_2014	0.238	0.238	0.197	0.197
	(0.633)	(0.633)	(0.516)	(0.516)
t_2012_2015	-0.725	-0.725	-0.737*	-0.737*
	(0.444)	(0.444)	(0.415)	(0.415)
t_2012_2016	-0.798**	-0.798**	-0.795**	-0.795**
	(0.395)	(0.395)	(0.380)	(0.380)
t_2012_2017	-1.336***	-1.336***	-1.319***	-1.319***
	(0.418)	(0.418)	(0.378)	(0.378)
t_2012_2018	-1.614***	-1.614***	-1.611***	-1.611***
	(0.311)	(0.311)	(0.300)	(0.300)
t_2012_2019	-1.612***	-1.612***	-1.619***	-1.619***
	(0.287)	(0.287)	(0.271)	(0.271)
t_2012_2020	-2.002***	-2.002***	-2.003***	-2.003***
	(0.399)	(0.399)	(0.384)	(0.384)

t_2012_2021	-1.520***	-1.520***	-1.538***	-1.538***
	(0.329)	(0.329)	(0.286)	(0.286)
t_2012_2022	-1.310***	-1.310***	-1.318***	-1.318***
	(0.175)	(0.175)	(0.158)	(0.158)
Observations	1092	1092	1092	1092

Note:* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses. Sample excludes years 2008, 2009, and 2010. Estimates calculated using Callaway & Sant'Anna (2021).

Appendix III. The Model: Callaway and Sant'Anna's (2021) Framework

Traditional DiD estimators yield unbiased estimates only under treatment effects that are homogeneous across units and time periods Baker and Larcker (2022), however, this assumption is unlikely to hold in the context of the EU ETS for three reasons: (1) The EU ETS was implemented in multiple phases with varying levels of stringency, (2) member states entered the scheme at different points in time, and (3) economic and energy structures differ substantially across countries. The staggered DiD approach addresses these challenges by estimating group–time average treatment effects ATT(*g, t*), capturing how policy impacts evolve for different adoption cohorts over time. Formally, this estimator is defined as follows:

(A3) $$ATT(g,t) = \mathbb{E}[Y_t(g) - Y_t(0) \mid G_i = g]$$

where $Y_t(0)$ represents the potential outcome without treatment, $Y_t(g)$ denotes the potential outcome if first treated in period g, and G_i indicates the treatment adoption period for country i. Finally, t is the time period in which the effect is evaluated. The Callaway and Sant'Anna (2021) estimator relies on several key assumptions, which are further assessed in Appendix I, to ensure the validity of the estimation: (1) Irreversibility of treatment: Once a country enters the EU ETS, it remains subject to the policy throughout the study period. This holds by design; no country has exited the system (except the UK via Brexit, which is addressed separately). (2) Cross-sectional sampling: There are no systematic correlations between countries beyond what the model captures. Because the dataset includes nearly the complete set of European countries participating in the EU ETS, along with non-participating countries used as control. Our estimates therefore characterize the policy effects within this specific set of countries rather than drawing inference about a broader population. Country inclusion is determined by geographic and policy criteria rather than emissions outcomes, which limits selection concerns. (3) No anticipation: Countries do not adjust behavior prior to formal policy adoption. This is plausible because compliance mechanisms only activate upon entry, firms typically delay costly adjustments until requirements

become binding, and pretreatment trends in emissions show no evidence of anticipatory effects.

(4) and (5) Parallel Trends: Untreated potential outcomes for treated and untreated groups would have followed parallel trends if the policy had not occurred. Two versions of this assumption apply depending on the specification used: i) Never-treated comparison is parallel trends between treated and never-treated countries, and ii) not-yet-treated comparison is parallel trends between treated and future-treated countries before treatment. Both approaches can be conditional on covariates. (6) Overlap: For each treated country, there must exist comparable controls in terms of pretreatment characteristics. This holds in our setting given the diversity of European energy systems and economic structures. This assumption is plausible because not all countries in the sample are fundamentally different. Even among non-ETS countries, there are likely several with pre-policy emission patterns, energy mixes, and economic profiles similar to those of ETS countries before treatment.

Identification

The staggered adoption of the EU ETS across member states created a quasi-experimental setting in which countries transitioned into the regulated regime at different points in time. Within this framework, each country–year observation is classified into one of three mutually exclusive states: (1) treated status indicating current participation in the EU ETS, (2) not-yet-treated status for countries that will enter in later periods, and (3) never-treated status for countries serving as permanent controls.

Formally, let $C = 1$ denote countries that never participated in the EU ETS during the study period (1990–2022). Let G represent the treatment adoption year when a country entered the trading system, and define D_t as a binary indicator equal to 1 when a country is subject to EU ETS regulations in year t. The outcome variable Y_t captures annual CO_2 emissions for each country in million tonnes ($MtCO_2$). Following the Callaway and Sant'Anna (2021) framework, we estimated the ATT parameters as $ATT(g, t)$, representing the estimated effect of EU ETS participation for countries that joined in year g, measured at time t. Under the proposed assumptions, this estimator yields consistent estimates of group-time average treatment effects, though we note that asymptotic properties should be interpreted with caution given the moderate number of cross-sectional units in our sample. We employed four different identification strategies to ensure robust identification of the treatment effects (following Callaway and Sant'Anna 2021):

- *Using never-treated countries as controls (for all $t \geq g$):*

This strategy provides a clean counterfactual by comparing treated units with those

permanently excluded from treatment as a control group. For all time periods $t \geq g$, where g represents the treatment adoption year and $C = 1$ denotes the set of countries that remain untreated, the ATT is identified as

(A4) $$ATT(g,t) = \mathbb{E}\left[Y_t - Y_{g-t} \mid G = g\right] - \mathbb{E}\left[Y_t - Y_{g-t} \mid C = 1\right]$$

- *Using not-yet-treated countries as controls (for all $t \geq g$):*

This estimator employs a dynamic control group where $D_t = 0$ indicates countries that have not yet received treatment at time t and $G = g$ specifies that the entry year into the system is not g. The ATT in this context is defined as

(A5) $$ATT(g,t) = \mathbb{E}\left[Y_t - Y_{g-t} \mid G = g\right] - \mathbb{E}\left[Y_t - Y_{g-t} \mid D_t = 0, G \neq g\right]$$

Both approaches initially rely on a basic staggered DiD estimator that captures the average difference in trends between treated and control groups over time. When covariates are introduced, the identification strategy is adjusted accordingly. Covariates not only allow for a more precise estimation by controlling for observed differences but also relax the parallel trends assumption. Instead of requiring that untreated potential outcomes follow parallel paths across all units, the assumption only needs to hold conditional on the included covariates.

- *Using covariates and never-treated countries as controls:*

The following equation implies that after conditioning on covariates X, the expected evolution of potential outcomes under nontreatment becomes identical between eventually treated and never-treated units:

(A6) $$\mathbb{E}[Y_t(0) - Y_{t-1}(0) \mid X, G = g] = \mathbb{E}[Y_t(0) - Y_{t-1}(0) \mid X, C = 1]$$

The ATT estimator is then implemented using a doubly robust method that combines inverse probability weighting and outcome regression. This approach is more reliable because it remains valid if at least one of these methods is correctly specified. Formally, this estimator adjusts for observable covariates X and is defined as the expected value of a weighted difference in outcomes in which the weights balance treated and control groups and the difference is adjusted for the estimated counterfactual outcomes.

(A7) $$\boldsymbol{ATT_{dr}^{nev}(g,t)} = \mathbb{E}\left(\frac{\boldsymbol{G_g}}{\mathbb{E}[\boldsymbol{G_g}]} - \frac{\frac{\boldsymbol{p_g(X)C}}{\boldsymbol{1-p_g(X)}}}{\mathbb{E}\left[\frac{\boldsymbol{p_g(X)C}}{\boldsymbol{1-p_g(X)}}\right]}\right)\left(\boldsymbol{Y - Y_{g-1} - m_{g,t}^{nev}(X)}\right)$$

where $m_{g,t}^{nev}(X) = \mathbb{E}\left[Y_t - Y_{g-1} \mid X, C = 1\right]$

and $p_g(X) = Pr\left(G_g = 1 \mid X, C = 1\right)$

where the first term in parentheses inside the expectation is the weight that compares treated units in group *g* to a weighted version of control units, using the generalized propensity scores $p_g(X) = P(G = g|X)$. Note that $\frac{G_g}{\mathbb{E}[G_g]}$ assigns weights to treated units in group *g* (*G* = 1 and *C* = 0) whereas the second term applies inverse-probability weights $\left(\frac{p_g(X)}{1-p_g(X)}\right) C$ to the control group (when *C* = 1 and *G* = 0). These weights are normalized to ensure balance. This weighting scheme has two key properties: for treated units (*G* = 1), the control term becomes 0, leaving only the direct treatment contribution. For control units (*C* = 1), the treatment term *G* becomes 0, retaining only the inverse-weighted control information. Further, note that the first term $\left(\frac{G_g}{\mathbb{E}[G_g]}\right)$ weights treated units equally whereas the second term $\left(\frac{\hat{p}^g(X)C}{1-\hat{p}^g(X)}\right)$ downweights control units unlikely to be treated and upweights those similar to treated units.

The second term is the outcome difference $Y_t - Y_{g-t}$ (the change in outcomes from the pretreatment to posttreatment period for treated countries) adjusted by subtracting an estimate of the counterfactual change $m_{g,t}^{nev}(X)$ for control units with the same covariates. Therefore, the estimator reweights control observations to make them comparable to treated observations based on covariates *X*. It adjusts the observed outcome differences by subtracting the predicted change that would have occurred for treated units if they had not been treated, based on similar control units. Significantly, the estimator combines the two methods using a doubly robust framework, which means the ATT estimate remains consistent if either the propensity score model or the outcome model is correctly specified, but not necessarily both.

- *Using covariates and not-yet-treated countries as control:*

The estimation framework employing not-yet-treated countries as controls follows a structure analogous to the never-treated specification, with key adjustments to account for the dynamic nature of the control group. The model retains the same doubly robust approach but replaces never-treated units with future adopters during their pretreatment periods. Formally, the equation can be expressed as:

(A8) $$\mathbb{E}[Y_t(0) - Y_{t-1}(0) \mid X, G = g] = \mathbb{E}[Y_t(0) - Y_{t-1}(0) \mid X, D_s = 0, G \neq g]$$

where D_t indicates whether a country is treated in a given year *t*.

For empirical estimation, the previous framework is used with subtle modifications to allow not-yet-treated countries to act as controls, but the interpretation remains the same.

(A9) $$\boldsymbol{ATT_{dr}^{ny}(g,t)} = \mathbb{E}\left(\frac{\boldsymbol{G_g}}{\mathbb{E}[\boldsymbol{G_g}]} - \frac{\frac{\boldsymbol{p_{g,t}(X)(1-D_t)}}{\boldsymbol{1-p_{g,t}(X)}}}{\mathbb{E}\left[\frac{\boldsymbol{p_{g,t}(X)(1-D_t)}}{\boldsymbol{1-p_{g,t}(X)}}\right]}\right)\left(\boldsymbol{Y - Y_{g-1} - m_{g,t}^{ny}(X)}\right)$$

$$\text{where } m_{g,t}^{ny}(X) = \mathbb{E}\left[Y_t - Y_{g-1} \,\middle|\, X, D_t = 0, G_g = 0\right]$$

$$\text{and } p_{g,t}(X) = Pr\left(G_g = 1 \,\middle|\, X, D_t = 0\right)$$

where $p_{g,t}(X)$ represents the conditional probability of treatment adoption in period *g*, given covariates *X* and not-yet-treated status; and $m_{g,t}^{ny}(X)$ estimates the expected outcome evolution for not-yet-treated units with characteristics *X*. The key difference is that now, instead of controls being $C = 1$ (never treated), the control group now includes $(1 - D_t)$, such that when countries are not treated ($D_t = 0$), they serve as controls.

The doubly robust estimator using the never-treated group compares the observed outcome evolution of treated units to that of units that are never treated, adjusting for covariates using inverse propensity weighting and outcome modelling. This approach is, as before, doubly robust and consistent if either the treatment assignment model or the outcome model is correctly specified.

Threats to Identification

Several potential threats to identification merit discussion. We address three concerns: violations of the stable unit treatment value assumption (SUTVA) through cross-border spillovers, confounding from EU accession and concurrent policies, and the endogeneity of carbon prices.

SUTVA and cross-border spillovers. A key identifying assumption is that a country's potential emissions depend only on its own treatment status, not on whether neighboring countries participate in the ETS. Because European energy markets show some degree of integration, this assumption warrants scrutiny. Cross-border spillovers could operate through two main channels: electricity trade and fuel price transmission.

Regarding electricity markets, during the sample period the electricity systems of the control countries (Albania, Bosnia and Herzegovina, Montenegro, North Macedonia, Serbia, Moldova, and Georgia) were only weakly integrated with the EU internal electricity market. Interconnection capacities between EU and non-EU systems were limited, and dispatch decisions were primarily determined by domestic conditions. This is consistent with EU policy targets: the European Commission's interconnection target of at least 15% of installed generation capacity was set for 2030, implying that effective integration was considerably lower during the study period and weaker still between EU and non-EU systems. OECD evidence confirms that cross-border electricity trade accounts for only a fraction of total consumption and that European electricity

markets remain partially fragmented, limiting the transmission of carbon cost signals across jurisdictions.

Regarding fuel markets, fossil fuel prices (natural gas, coal, oil) are determined in global or continental commodity markets. The EU ETS, while the largest carbon market, covers approximately 40% of EU emissions and is too small relative to global fossil fuel markets to materially alter world prices for coal or natural gas. Price formation for these commodities is driven by global supply and demand conditions, geopolitical factors, and weather patterns, not by the carbon cost imposed on European generators. Consequently, the scope for SUTVA violations through fuel price channels is limited.

Crucially, to the extent that spillovers nevertheless occur, they would bias our estimates toward zero rather than generate spurious effects. Carbon pricing under the ETS raises marginal generation costs in regulated countries, which strengthens incentives to import energy from unregulated neighbors. This would increase production and emissions in the control group, reducing the treatment–control differential. Our estimates should therefore be interpreted as conservative lower bounds of the policy's true effect. We also note that the Carbon Border Adjustment Mechanism (CBAM), which could affect cross-border competitive dynamics, only entered its transitional phase in 2023 and will not be fully implemented until 2026—well after the end of our sample period.

EU accession and confounding policies. For several countries in our sample, the timing of ETS entry coincides with EU accession, raising the concern that estimated effects may reflect the broader suite of EU environmental directives rather than the ETS specifically. We address this concern with three arguments. First, EU accession is a gradual and highly structured process requiring progressive adoption of the acquis communautaire over many years. Environmental legislation is one of the most demanding chapters of accession, requiring extensive transposition well before formal membership. Regulatory convergence therefore occurs progressively during the pre-accession period, not as a discrete policy shock at the accession date. If EU membership—rather than the ETS—were driving the results, we would expect to observe differential pre-treatment trends during the convergence process. We find no evidence of such differential trends in either unconditional or conditional specifications.

Second, some ETS participants entered without simultaneously joining the EU. Norway and Iceland joined the ETS in 2008 through EEA agreements, not via EU accession. If EU membership were the true treatment, we would not expect emissions reductions for these non-EU ETS participants. Our estimates for the 2008 cohort, which includes Norway and Iceland, show treatment effects consistent with those of EU member states, providing direct evidence that

the ETS generates abatement effects independent of EU accession.

Third, the specific concurrent policies cited as potential confounders do not introduce additional binding constraints on our outcome variable. The Industrial Emissions Directive (IED) regulates industrial emissions primarily through technology-based performance standards and explicitly excludes CO_2 emissions that fall under the scope of the ETS. Since our analysis focuses exclusively on CO_2 emissions from the energy sector—activities comprehensively regulated under the ETS—the IED does not impose additional constraints on the outcome of interest. The EU Climate and Energy Package (CEP) combines ETS pricing with renewable energy and energy efficiency targets, but these complementary instruments primarily affect non-ETS sectors or operate indirectly. We include renewable energy share as a covariate in conditional specifications to account for structural energy transition trends. Finally, in countries with pre-existing national carbon pricing policies (notably Sweden, Finland, and Denmark), the introduction of the ETS largely replaced overlapping national instruments rather than introducing an entirely new regulatory constraint. This substitution reduces policy discontinuity at the treatment date and attenuates rather than inflates estimated effects.

Carbon price variation. ETS allowance prices varied substantially across phases, from near-zero during Phase I to sustained levels above €20 in Phase III. While this variation is informative for understanding the ETS, we do not include carbon prices directly as a time-varying covariate because they are determined endogenously within the system and jointly reflect policy design choices (cap stringency, allocation rules) and realized emissions reductions. Conditioning on an endogenous intermediate variable would attenuate or distort the estimated policy effects. Instead, our staggered design captures variation in policy stringency through phase-specific treatment effects, which implicitly reflect the price dynamics associated with each phase's design features.

Estimators: Group-Time Average Treatment Effects Aggregation

Due to the nature of the panel data, many years and few groups lead to many disaggregated group-time average treatment effects. That is why there are some benefits from aggregating them in different ways (Callaway and Sant'Anna 2021). This provides a straightforward approach to visualizing heterogeneous effects over time. We used four aggregate estimates. First was the simple average ATT that aggregates all available group–time estimates into a single parameter, assigning equal weight to each group–period cohort. This provides a global measure of the EU ETS impact across all phases and periods, where *T* denotes the final time period.

(A10) $$ATT^{simple} = \frac{1}{\sum_{g \epsilon \mathcal{G}} (T-g+1)} \sum_{g \epsilon \mathcal{G}} \sum_{t=g}^{T} ATT(g,t)$$

For the evaluation of the ATT disaggregated by group, ATT^{cohort} represents a clear and

interpretable parameter. It captures the average effect for the adoption cohort *g* across all posttreatment periods, with each period weighted equally. The term $(T - g + 1)$ reflects the number of posttreatment periods, spanning the period from the adoption year *g* to the final observation period *T*.

(A11) $$ATT^{cohort}(g) = \frac{1}{T-g+1}\sum_{t=2}^{T} 1\{g \leq t\}\, ATT(g,t)$$

To examine the effect evolution across calendar time, this time-specific ATT shows how treatment effects vary over time across all groups treated by time *t*, weighted for the number of units treated in each group–time cell and giving preference to bigger cohorts.

(A12) $$ATT^{calendar}(t) = \sum_{g \epsilon \mathcal{G}} 1\{t \geq g\} P(G = g | G \leq t)\, ATT(g,t)$$

Finally, to estimate the dynamic effects relative to treatment timing, the event study ATT^{event} estimates the average treatment effect at *e* periods after adoption, properly weighted by cohort size.

(A13) $$ATT^{event}(e) = \sum_{g=2}^{T} 1\{g + e \leq T\}\, ATT(g, g+e)\; P(G = g | G + e \leq T)$$

where $ATT(g, g+e)$ is the group-time effect for cohort g at event time *e* (i.e., relative period). The indicator $1\{g + e \leq T\}$ ensures that only feasible (g, e) combinations are used. The term $P(G = g \mid G + e \leq T)$ represents the weight that accounts for the probability of belonging to cohort g conditional on having observed data at event time *e*. Finally, it is important to note that all weights correspond to the conditional probabilities $P(G = g \mid G \leq t)$ specified in the framework of Callaway and Sant'Anna (2021).

Empirical Strategy

To model the EU ETS, we divided the analysis into four phases corresponding to each country's entry into the system: Phase 1 (2005–2007), Phase 2 (2008–2012), Phase 3 (2013–2020), and Phase 4 (2021–2030). Because data are only available through 2022, results for the final phase are limited to that year (see the data description in the following section). We implemented this division into phases using the *csdid* package in STATA (Sant'Anna and Zhao 2020). By default, robust and asymptotic standard errors clustered at country level were estimated using the doubly robust method (dr). Finally, renewable-energy share was included as the sole covariate.

Outcome Regression Estimation

When no covariates are included, the estimator reduces to the standard staggered DiD design; it compares the before–after change in emissions for treated units with the corresponding change for control units. Under unconditional parallel trends, an assumption shown to hold later, the average treatment effect on the treated (ATT) is simply the raw difference in emissions trends between treated and untreated countries. No explicit counterfactual predictions are required. With

covariates X (here, the renewable-energy share), the outcome-regression component of *csdid* predicts counterfactual trends conditional on X. In this setting the model actively estimates how emissions would have evolved for treated units had they not entered the EU ETS:

(A14) $$m_{g,t}(X) = \mathbb{E}\left[Y_t(0) - Y_{g-1}(0) \middle| X, \text{Control group}\right]$$

where $m_{g,t}(X)$ is the predicted counterfactual change for a treated unit observed in the posttreatment period t (first treated in period g), given its covariates X. Practically, the control group's emissions trend, adjusted for covariates, serves as the benchmark for the treated group. Then $m_{g,t}(X)$ is estimated via OLS regression using control units (never-treated or not-yet-treated), specified as follows:

(A15) $$Y_t - Y_{g-1} = \hat{\beta}_0 + \hat{\beta}_1 renewable\_share + \varepsilon$$

Propensity Score Estimation:

The propensity score is estimated only in specifications that include covariates because it captures the probability that a country first receives treatment in period g conditional on its pretreatment characteristics X. Formally,

(A16) $$p_g(X) = Pr(G = g \mid X, G \geq g)$$

Following Callaway and Sant'Anna (2021), $p_g(X)$ is estimated by inverse-probability weighting, fitting a logistic regression of the treatment indicator on the covariates that describe country characteristics. These weights help balance the covariate distribution between treated and control countries because treated countries are typically larger than the never-treated controls.