Quantifying the Effects of Climate Policy Stringency on Verified Emissions and Satellite-Derived NOx

Master's Thesis

Arnav Agrawal

MA in Quantitative Methods in the Social Sciences, Columbia University.

Abstract

I investigate how European Union Emissions Trading System (EU ETS) policy stringency affects both verified CO_2 emissions and satellite-derived NOx proxies at 521 large combustion plants across Europe (2018–2023). Using two-way fixed effects with facility and region-by-year fixed effects, I find that a 10% allocation shortfall is associated with 1.9% lower verified CO_2 emissions ($\boldsymbol{\beta = -0.186}$, $\boldsymbol{p < 0.001}$). Effects are strongest for coal-dominant facilities ($\boldsymbol{\beta = -0.98}$) and electricity generators ($\boldsymbol{\beta = -0.21}$). Satellite-derived NOx—constructed via Beirle-style flux-divergence from TROPOMI observations—shows consistent negative effects at conservative detection limits ($\boldsymbol{\beta = -0.003}$, $\boldsymbol{p < 0.01}$), providing independent corroboration. The framework makes three methodological contributions: (i) geospatial foundation model embeddings (Google AlphaEarth) as high-dimensional controls for satellite retrieval confounders; (ii) NUTS2 regional clustering for inference with PyPSA-Eur power system clusters for electricity-sector heterogeneity analysis; and (iii) a simplified Beirle-style NOx quantification method adapted for panel econometrics. The dual-outcome approach enables cross-validation: directional agreement between administrative CO_2 and satellite NOx strengthens confidence that both capture genuine policy effects. This work demonstrates how ML-derived geospatial features and satellite remote sensing derived physical observables can be integrated into causal inference frameworks for econometric analysis of point-source emissions.

Keywords: Climate Policy, Climate Monitoring, EU ETS, Verified Emissions, Satellite Remote Sensing, TROPOMI, NOx Emissions, Flux Divergence, Difference-in-Differences, Causal Inference, Large Combustion Plants

1 Introduction

Evaluating climate policy requires measuring actual emission outcomes. The European Union Emissions Trading System (EU ETS) generates rich administrative data on verified CO_2 emissions at the installation level, providing the gold standard for measuring greenhouse gas output from regulated facilities. However, relying solely on self-reported emissions raises questions about verification and leaves unmeasured the local air quality co-benefits that accompany carbon reductions. Satellite remote sensing offers an independent, physically-grounded approach to quantifying emissions from space, potentially revealing both verification opportunities and co-pollutant dynamics that administrative data cannot capture.

This study adopts a dual-outcome approach that exploits the complementary strengths of administrative and satellite data. The two outcomes are: (i) verified EU ETS CO_2 emissions—high-quality, installation-level measures from the EU ETS registry that provide accurate compliance trajectories and absolute emission levels; and (ii) a satellite-derived NOx emission proxy—a physically grounded indicator constructed from TROPOMI NO_2 tropospheric columns and ERA5 winds, following the flux-divergence approach of Beirle et al. [1–3].

Why use both outcomes? CO_2 is a well-mixed greenhouse gas with global climate impacts; nitrogen oxides (NO_x), by contrast, are criteria pollutants whose health effects—respiratory illness, cardiovascular disease, premature mortality—fall disproportionately on populations living near emission sources. As [4] emphasize, air quality co-benefits are particularly policy-relevant because they are local and immediate, whereas averted climate damages are global and long-term. The dual-outcome design provides: (i) verified emissions for accurate policy effect estimation, (ii) satellite-derived NOx for testing co-benefit hypotheses, and (iii) cross-validation opportunities where both outcomes should respond to common policy shocks.

This study develops a novel framework for evaluating climate policy impacts using both administrative emissions data and satellite remote sensing. I focus on the European Union Emissions Trading System (EU ETS), the world's largest carbon market, which creates economic incentives for industrial facilities to reduce CO_2 emissions through a cap-and-trade mechanism. The framework addresses two fundamental methodological challenges: (i) constructing a satellite-derived NOx emission proxy that is physically interpretable and appropriate for panel econometric analysis, and (ii) controlling for high-dimensional confounders that affect both policy exposure and emission outcomes.

The study makes three methodological contributions. The first two follow a recent trend in causal inference toward incorporating machine learning-derived features to address high-dimensional confounding [5–7]; the third adapts an atmospheric physics method for panel econometrics.

First, I demonstrate the use of geospatial foundation model embeddings as high-dimensional controls in panel-based climate monitoring studies. Specifically, I incorporate Google AlphaEarth embeddings [8]—64-dimensional representations learned from multi-source satellite imagery, climate reanalysis, and geotagged text—as control variables in the econometric specifications. These embeddings capture between-unit heterogeneity arising from local geographic, infrastructural, and climate

context in a data-efficient manner, providing a scalable approach to controlling for high-dimensional spatial confounders that would be impractical to specify manually. This application extends prior work on learned representations for causal inference—originally developed for text embeddings [6]—to the domain of geospatial environmental monitoring. The approach is particularly suited to difference-in-differences settings where high-dimensional confounders may violate the parallel trends assumption if left uncontrolled [7].

Second, I use Eurostat NUTS2 regions for spatial clustering in both fixed effects structure and inference. Standard errors are clustered by NUTS2 region, which groups facilities that share common regional economic conditions, labor markets, and policy enforcement mechanisms. The same regions define Region×Year fixed effects, absorbing time-varying regional confounders that correlate with both policy exposure and air quality outcomes. Unlike sector-specific clustering (e.g., power system network clusters), NUTS2 regions apply uniformly to all industrial facility types and correspond to administrative units where EU and national environmental policies are implemented. For electricity sector heterogeneity analysis, I additionally implement PyPSA-Eur power system clusters [9]—k-means clusters computed on transmission network topology—which group facilities facing correlated wholesale prices and dispatch patterns.

Third, for the satellite outcome, I implement a simplified Beirle-style flux-divergence method that provides physically grounded NOx emission estimates at the facility level. The approach computes the advection—the scalar product of wind velocity and the spatial gradient of NO_2 column density—which under the continuity equation is proportional to local emissions minus chemical loss. For each facility, I integrate advection over a 15 km disc, apply a lifetime correction following [3], and convert to NOx emission rates. This methodology follows the Beirle et al. (2019, 2021, 2023) family of methods [1–3], which are physically transparent, computationally tractable for known point sources, and specifically designed for power-plant-scale NOx plumes.

The analysis links three data sources on large combustion plants (LCPs) in the EU: (i) the European Environment Agency's LCP registry providing plant characteristics and coordinates, (ii) EU ETS compliance data providing installation-level verified emissions and free allocations, and (iii) TROPOMI satellite observations processed through the Beirle-style flux-divergence methodology to derive NOx emission proxies. Policy exposure is measured continuously through the *allocation ratio*—free allowances divided by verified emissions—where values below unity indicate facilities must purchase additional permits, creating direct economic pressure to reduce emissions.

The econometric framework employs two-way fixed effects (TWFE) specifications with facility and region×year fixed effects. The Callaway-Sant'Anna estimator [10] proved infeasible due to panel structure (Section 2.4).

By demonstrating that both administrative emissions records and satellite-derived NOx estimates can provide individual-emitter-level, policy-parameterized estimates of emission responses to carbon market stringency, this work contributes to the emerging literature on comprehensive climate policy evaluation. The dual-outcome

approach enables testing whether policy effects on verified CO_2 are accompanied by corresponding changes in satellite-observed combustion co-pollutants.

2 Background and Literature Review

2.1 The EU Emissions Trading System

The EU ETS, established in 2005, operates as a cap-and-trade system covering approximately 40% of EU greenhouse gas emissions. Large combustion plants with thermal input exceeding 20 MW are required to hold European Union Allowances (EUAs) equal to their verified CO_2 emissions. Allowances are distributed through a combination of free allocation (based on historical benchmarks and carbon leakage risk) and auctioning. Installations that emit more than their free allocation must purchase additional allowances, creating marginal abatement incentives [11].

The policy has evolved through fmy phases, with Phase III (2013–2020) and Phase IV (2021–2030) introducing progressively tighter caps and reduced free allocation, particularly for the power sector. This study focuses on the period 2018–2023, spanning the transition from Phase III to Phase IV and capturing significant variation in policy stringency across facilities.

2.2 Satellite-Based Air Quality Monitoring

The TROPOMI instrument aboard Sentinel-5P, operational since late 2017, provides daily global observations of tropospheric NO_2 column densities at unprecedented spatial resolution ($\sim$3.5$\times$5.5 km^2 at nadir). This represents a significant improvement over predecessor instruments (OMI, GOME-2) and enables detection and quantification of emissions from individual point sources [3, 12].

Previous studies have used satellite observations to verify emission reductions from policy interventions. [13] demonstrated that China's ultra-low-emission retrofits for coal-fired power plants produced measurable NO_2 declines visible from space. [14] documented substantial NO_x reductions over Europe between 1996 and 2010, attributing these to environmental policies and economic recession. However, these studies typically analyze aggregate regional trends rather than plant-level responses to specific policy parameters.

2.3 Satellite-Based NOx Emission Quantification: The Flux-Divergence Approach

A key methodological challenge in quantifying emissions from satellite-observed NO_2 is separating the source signal from background concentrations and converting column densities to emission rates. This challenge is particularly acute in Europe, where high population density means that most large combustion plants are located in or near urban areas, surrounded by other pollution sources (traffic, industry, heating).

The flux-divergence (or advection) approach, developed by Beirle et al. [1–3], provides a physically grounded solution. The method exploits the continuity equation: horizontal NO_2 fluxes $\mathbf{F} = \mathbf{w}V$ (where $\mathbf{w}$ is wind velocity and V is tropospheric

vertical column density) satisfy

$$\nabla \cdot \mathbf{F} = E - S \tag{1}$$

where E represents local emissions and S represents chemical sinks. Under typical conditions where wind field divergence is negligible, this reduces to the advection formulation:

$$A = \mathbf{w} \cdot \nabla V \approx E - S \tag{2}$$

The advection A measures the downwind rate of change in NO_2 column density and is particularly sensitive to strong point sources, which create sharp spatial gradients in the NO_2 field.

Beirle et al. (2021) [2] presented the first global catalog of NOx point source emissions derived from TROPOMI using this approach, identifying 451 sources. Beirle et al. (2023) [3] introduced version 2 with several improvements: use of the PAL (Products Algorithm Laboratory) NO_2 product with higher column densities (factor of 1.1–1.4), corrections for plume height effects on satellite sensitivity, topographic corrections, and a lifetime correction to account for chemical loss within the integration radius. These refinements resulted in emission estimates approximately 3 times higher than version 1, with validation showing agreement within 20% of reported emissions from the German Environment Agency (UBA) and US EPA.

[15] developed an alternative regression-based approach for decomposing TROPOMI NO_2 into urban, industrial, and background components during COVID-19, demonstrating that wind information can isolate individual source contributions even in complex emission environments. [12] extended the methodology to megacities, estimating both emissions and effective NOx lifetimes through simultaneous fitting of downwind plume evolution.

This study adopts the Beirle family of methods because they are: (i) physically transparent, grounded in the continuity equation; (ii) computationally tractable for known point sources; and (iii) specifically designed and validated for power-plant-scale NOx plumes. I implement a simplified version appropriate for panel econometric analysis, acknowledging the additional uncertainty from using OFFL L3 data rather than the PAL product.

2.4 Causal Inference with Staggered Treatment Timing

Standard two-way fixed effects estimators can produce biased estimates when treatment timing varies across units and treatment effects are heterogeneous [16]. Recent methodological advances, including the Callaway and Sant'Anna [10] and Sun and Abraham [17] estimators, address these concerns by constructing treatment effect estimates using only valid comparisons (treated versus not-yet-treated or never-treated units) and allowing for treatment effect heterogeneity across cohorts and time.

I initially planned to complement TWFE with the Callaway-Sant'Anna estimator. However, this approach proved infeasible: of 457 ever-treated facilities (defining treatment as $R_{it} < 1$), 386 (84.5%) were already treated in 2018—the first year of my panel. With no pre-treatment observations for 84.5% of treated units, the estimator cannot form a meaningful untreated counterfactual cohort.

I therefore interpret β_{TWFE} as an *average response to changes in* R_{it}, not as a staggered treatment timing effect. I do **not** interpret the estimates as event-study dynamics; pre-trend testing is left to future work with an extended 2013–2017 panel (EU ETS Phase 3). The continuous treatment approach exploits within-facility variation in policy stringency over time and remains valid under homogeneous treatment effects, which the heterogeneity analysis (Section 5.3) examines empirically.

2.5 High-Dimensional Controls and ML-Derived Features in Causal Inference

A growing literature in causal inference addresses the challenge of controlling for high-dimensional confounders—settings where the number of potential control variables is large relative to sample size, or where relevant confounders are difficult to specify manually. The foundational work of [5] established the "double/debiased machine learning" framework, showing how machine learning methods can be used to estimate nuisance parameters (propensity scores, outcome regressions) while maintaining valid inference on treatment effects. This approach enables researchers to control for high-dimensional confounders without imposing restrictive parametric assumptions.

In the difference-in-differences context specifically, [7] developed efficient estimators for settings where the parallel trends assumption holds only conditional on high-dimensional covariates. This is particularly relevant when unobserved confounders that violate parallel trends can be proxied by high-dimensional observables—such as detailed geographic or economic characteristics that would be impractical to specify manually but can be captured through flexible ML methods.

A parallel development concerns the use of *learned representations*—embeddings from neural networks or foundation models—as control variables. [6] demonstrated that text embeddings can serve as effective controls for confounding in observational studies, provided the embeddings capture the relevant confounding information. The key insight is that pre-trained (usually unsupervised or self-supervised) representations encode information about latent confounders that would otherwise be unobserved. This approach has been extended to various domains, including image embeddings and, most recently, geospatial foundation models.

For clustered inference, [18] established theoretical foundations for network cluster-robust standard errors. They show that valid cluster-robust inference requires clusters with low "conductance"—the ratio of edges crossing cluster boundaries to total edges within clusters. This implies that clusters should be defined based on the correlation structure of the data-generating process, not arbitrary geographic or administrative boundaries. When observations are connected through a network (as power plants are through the transmission grid), clusters derived from network topology can satisfy these requirements.

This study contributes to this literature by demonstrating two novel applications: (i) using geospatial foundation model embeddings (AlphaEarth) as controls for spatial confounding in environmental panel data, and (ii) using k-means clusters derived from power system network features (PyPSA-Eur) for heterogeneity analysis. To my knowledge, this represents the first application of model-derived clustering—where clusters

are computed on features from an external domain-specific model rather than on the outcome data itself—for econometric time series analysis.

3 Data

This section describes the data sources, processing pipeline, and construction of the analysis panel. The study combines administrative records on industrial facilities and EU ETS compliance with satellite remote sensing and meteorological reanalysis data.

3.1 Data sources

3.1.1 EEA Large Combustion Plant Registry

The European Environment Agency (EEA) maintains the Industrial Emissions Portal, which includes the Large Combustion Plant (LCP) dataset. This registry provides annual reports on combustion plants with rated thermal input $\geq$50 MW, including:

- Geographic coordinates (latitude, longitude)
- Plant identification (LCP INSPIRE ID, installation name)
- Rated thermal capacity (MW)
- Annual fuel consumption by fuel type (TJ)
- Country of operation

The raw dataset contains 3,405 unique plant parts for the period 2018–2023. After filtering for complete capacity and fuel data, 2,821 plants remain with valid time-varying attributes.

3.1.2 EU ETS Compliance Data

EU ETS installation-level compliance data is obtained from the European Union Transaction Log (EUTL), accessed via the `pyeutl` Python package. For each installation-year, the data includes:

- Verified CO_2 emissions (tCO_2)
- Free allocation of allowances (tCO_2-equivalent)
- Surrendered allowances (tCO_2-equivalent)
- Installation identifier and country

The LCP and ETS datasets are linked through the EU Registry on Industrial Sites, which provides crosswalk tables mapping LCP installation parts to their parent ETS installations via normalized identifiers.

3.1.3 TROPOMI Satellite Observations

Tropospheric NO_2 column densities are obtained from the Sentinel-5P TROPOMI instrument via Google Earth Engine, using the OFFL (offline) L3 product (`COPERNICUS/S5P_OFFL_L3_NO2`). TROPOMI provides daily global coverage at approximately 3.5×5.5 km^2 spatial resolution. Quality-filtered observations are used,

retaining only pixels with quality assurance values ≥0.75. TROPOMI captures approximately 14 orbits per day globally, with each orbit covering a distinct swath (∼2600 km); for any given facility, only one orbit per day provides valid coverage.

Importantly, Beirle et al. (2023) [3] use the PAL (Products Algorithm Laboratory) NO_2 product, which provides higher tropospheric vertical column densities (TVCDs) than the OFFL product by a factor of approximately 1.1–1.4, due to updated retrieval algorithms and air mass factor corrections. This difference, combined with other methodological refinements, contributed to their version 2 emission estimates being approximately 3 times higher than version 1. Since I use the OFFL L3 product available via Google Earth Engine rather than the PAL product, the satellite-derived NOx estimates carry additional uncertainty (approximately ±25% relative to PAL-based estimates) that must be acknowledged.

3.1.4 ERA5-Land Reanalysis

Hmyly 10-meter wind components (u_{10}, v_{10}) are obtained from the ERA5-Land reanalysis product via Google Earth Engine. Daily mean wind speed and direction are computed at each facility location for the advection calculation. Following Beirle et al. [3], days with wind speeds below 2 m/s are excluded, as weak winds produce unreliable advection estimates and allow plumes to stagnate near sources. Additionally, observations where the lifetime correction factor $c_\tau \geq 3$ are dropped, as this exceeds the typical range of 1.2–1.8 reported by Beirle et al. Facility-years with fewer than 20 valid observation days (after wind filtering) are excluded from the satellite panel, as statistical uncertainty becomes prohibitively large with insufficient temporal sampling.

3.1.5 SRTM Digital Elevation Model (DEM)

The Shuttle Radar Topography Mission provides a near-global digital elevation model at 1 arc-second (30 m) resolution. I use the USGS/SRTMGL1_003 dataset on Google Earth Engine [19, 20] to compute surface elevation gradients used in the topographic correction (Eq. 15).

3.1.6 Urbanization Classification

Facilities located within urban areas experience higher background NO_2 concentrations from traffic and other distributed sources, which adds noise to satellite-derived emission estimates. To enable heterogeneity analysis and descriptive statistics, each facility is assigned an urbanization degree from the JRC Global Human Settlement Layer Degree of Urbanisation raster (GHS-SMOD R2023A) [21]. The SMOD classification ranges from 10 (water) through rural categories (11–13) to suburban (21) and urban categories (22–30), based on population density and built-up area from satellite imagery.

Two urbanization variables are constructed:

- `urbanization_degree`: The continuous SMOD code (10–30) at each facility location
- `in_urban_area`: A boolean flag indicating SMOD ≥ 21 (suburban or denser)

These variables are collected for heterogeneity analysis (comparing treatment effects across urban vs. rural subsamples) and descriptive statistics, *not* as regression controls. The AlphaEarth embeddings (64 dimensions) already encode land use, built-up area, and urbanization patterns implicitly. Including an explicit urbanization control would introduce multicollinearity with the embedding dimensions without improving identification, since urbanization is time-invariant and absorbed by facility fixed effects regardless. The proper causal use of urbanization is for split-sample analysis, not as an additional covariate.

3.2 Facility Construction: Spatial Clustering

Individual LCP plant parts may represent components of larger industrial complexes. To avoid treating co-located plants as independent units, I apply spatial clustering using a 500-meter threshold. Plants within 500m of each other are grouped into a single *facility* using a union-find algorithm.

Let $\mathcal{P} = \{p_1, \ldots, p_N\}$ denote the set of LCP plants with coordinates (ϕ_j, λ_j) for plant j. The distance between plants j and k is computed using the WGS84 ellipsoidal approximation [22]:

$$d_{jk} \approx \sqrt{(m_\phi \cdot \Delta\phi_{jk})^2 + (m_\lambda \cdot \Delta\lambda_{jk})^2} \tag{3}$$

where the latitude scale factor follows the WGS84 series expansion:

$$m_\phi = 111{,}132.954 - 559.822\cos(2\bar{\phi}) + 1.175\cos(4\bar{\phi}) \quad \text{[m/deg]} \tag{4}$$

and the longitude scale factor varies with latitude:

$$m_\lambda = 111{,}132.954 \times \cos(\bar{\phi}) \quad \text{[m/deg]} \tag{5}$$

where $\bar{\phi}$ is the mean latitude of the dataset. The latitude formula is accurate to 0.01 m per degree; the longitude formula has <0.5% error compared to the full WGS84 ellipsoidal expression. This precision is more than sufficient for identifying co-located plants, as the 500m clustering threshold is conservative relative to the spatial extent of most industrial complexes.

Plants are grouped into facility i if they form a connected component under the relation $d_{jk} < 500$m. For each facility, the centroid coordinates are computed as the arithmetic mean of constituent plant coordinates:

$$(\bar{\phi}_i, \bar{\lambda}_i) = \frac{1}{|F_i|} \sum_{j \in F_i} (\phi_j, \lambda_j) \tag{6}$$

where F_i denotes the set of plants in facility i.

This clustering reduces the sample from 1,576 individual plants with ETS linkage to 932 facilities, of which 318 are multi-plant facilities.

3.3 Time-Varying Attributes

3.3.1 Capacity and Fuel Shares

For each facility-year (i, t), rated thermal capacity is aggregated as the sum across constituent plants:

$$\text{Capacity}_{it} = \sum_{j \in F_i} \text{Capacity}_{jt} \quad [\text{MW}] \tag{7}$$

Fuel energy consumption is similarly aggregated, then converted to fuel shares. Let $E_{it}^{(f)}$ denote total energy consumption from fuel type $f \in \{\text{gas}, \text{coal}, \text{oil}, \text{biomass}, \text{other}\}$ for facility i in year t, measured in terajoules (TJ). Fuel shares are computed as:

$$s_{it}^{(f)} = \frac{E_{it}^{(f)}}{\sum_{f'} E_{it}^{(f')}} \tag{8}$$

Fuel types used by fewer than 10% of facility-years (Other Solid, Peat) are dropped, shares renormalized, and facilities with no remaining fuel coverage are excluded from the sample (although there were no such facilities).

3.3.2 ETS Policy Exposure Variables

The key treatment variable is the *allocation ratio*, defined as:

$$R_{it} = \frac{A_{it}}{V_{it}} \tag{9}$$

where A_{it} is total free allocation and V_{it} is verified emissions for facility i in year t, both in tCO_2. Values $R_{it} < 1$ indicate the facility must purchase additional allowances on the carbon market, representing increased policy stringency.

The *shortfall* is defined as:

$$S_{it} = V_{it} - A_{it} \tag{10}$$

Positive shortfall indicates emissions exceed free allocation.

Facilities with allocation ratios outside the range $[0.01, 20]$ are excluded as likely data errors or non-operating installations.

3.4 Satellite NOx Emission Proxy: Beirle-Style Flux-Divergence

The satellite outcome variable is constructed using a simplified Beirle-style flux-divergence method, following the approach developed by Beirle et al. [1–3]. This method provides physically grounded NOx emission estimates by exploiting the relationship between wind-driven advection and local emissions.

3.4.1 Identification versus Quantification

Beirle et al.'s v2 catalog combines two distinct algorithmic components: (i) an automatic point-source *identification* algorithm that locates emission maxima in the global

advection field, and (ii) a *quantification* method that estimates emission rates by spatially integrating advection around each identified source. Crucially, the authors note that "the quantification of NOx emissions by spatial integration of the corrected advection map could be applied to these locations or *any other known point source*" [3].

In this study, I *skip the identification step* because I already have a curated set of ETS/LCP facilities with reliable coordinates from the European Environment Agency registry. I apply Beirle's quantification method directly to these known source locations.

To guard against treating noise as signal, I implement *simplified significance flags* that parallel Beirle's catalog selection criteria:

- **Detection limit**: Emission estimates below 0.11 kg/s are flagged, corresponding to Beirle's standard detection threshold for non-desert conditions.
- **Statistical integration error**: Facilities with >30% relative statistical uncertainty in the spatial integration are flagged.
- **Spatial interference**: Facilities with another ETS facility within 20 km are flagged, as their satellite outcome may reflect cluster-level rather than single-facility emissions.

These flags are used in sensitivity analyses rather than for hard filtering, preserving the full panel while allowing transparent restriction to "significant" satellite observations.

3.4.2 Advection Formulation

The advection A is defined as the scalar product of wind velocity and the spatial gradient of NO_2 tropospheric vertical column density (TVCD):

$$A = \mathbf{w} \cdot \nabla V = u\frac{\partial V}{\partial x} + v\frac{\partial V}{\partial y} \tag{11}$$

where $\mathbf{w} = (u, v)$ is the horizontal wind vector (m/s) from ERA5-Land and V is the NO_2 TVCD (molec/m^2). Under the continuity equation, this advection is proportional to local emissions minus chemical sinks.

For each facility i and day d, spatial gradients are computed on a local grid (30 km $\times$ 30 km centered on the facility) using finite differences on the TROPOMI L3 lat–lon grid:

$$\frac{\partial V}{\partial x} \approx \frac{V(x + \Delta x, y) - V(x - \Delta x, y)}{2\Delta x} \tag{12}$$

$$\frac{\partial V}{\partial y} \approx \frac{V(x, y + \Delta y) - V(x, y - \Delta y)}{2\Delta y} \tag{13}$$

where Δx and Δy correspond to the TROPOMI grid resolution (approximately 3.5 km $\times$ 5.5 km). This differs from Beirle et al., who compute derivatives on the native TROPOMI pixel grid to handle cloud-induced gaps; the L3 gridded product used here introduces additional smoothing and potential artifacts.

3.4.3 NO_2 to NOx Scaling

TROPOMI measures NO_2, but NOx emissions include both NO and NO_2. Following Beirle et al. [3], I apply a scaling factor c_{NOx} based on the photostationary state (PSS):

$$c_{\mathrm{NOx}} = \frac{[\mathrm{NOx}]}{[\mathrm{NO_2}]} = 1 + \frac{J}{k[\mathrm{O_3}]} \tag{14}$$

where J is the NO_2 photolysis frequency (parameterized as $0.0167 \times \exp(-0.575/\cos(\mathrm{SZA}))\mathrm{s}^{-1}$, where SZA is the Solar Zenith Angle of the observation), k is the reaction rate constant for NO + O_3 ($2.07 \times 10^{-12} \times \exp(-1400/T)$ cm^3 molec^{-1} s^{-1}), and $[O_3]$ is taken from an ozone climatology. For detected point sources, Beirle et al. report a typical NOx/NO_2 ratio of approximately 1.38 ± 0.10.

Following Beirle et al., I apply a fixed scaling factor of $c_{\mathrm{NOx}} = 1.38$ with uncertainty ± 0.10 (approximately 7% relative uncertainty), which represents the empirically observed mean ratio across detected point sources.

3.4.4 Topographic Correction

Over mountainous terrain, 3D radiative transfer effects cause systematic artifacts in the advection field. Following Beirle et al. [3] Sect. 3.7, I apply a topographic correction:

$$A^* = A + f \cdot C_{\mathrm{topo}}, \quad C_{\mathrm{topo}} = \frac{V}{H_{\mathrm{sh}}} \cdot (\mathbf{w}_0 \cdot \nabla z_0) \tag{15}$$

where V is the NO_2 TVCD, $H_{\mathrm{sh}} = 1$ km is the assumed NOx scale height, $\mathbf{w}_0 \cdot \nabla z_0$ is the dot product of the surface wind vector and the surface elevation gradient (from the SRTM Digital Elevation Model (DEM) [19, 20] via Google Earth Engine), and $f = 1.5$ is an empirically derived scaling factor (Appendix A of Beirle et al.). The combined effect yields an effective scale height of $1/1.5 = 667$ m. For flat terrain typical of European power plant locations, this correction is small.

3.4.5 Spatial Integration and Lifetime Correction

For each facility, the raw emission rate is computed by spatially integrating the topography-corrected advection A^* over a 15 km disc around the facility location (Beirle Eq. 11):

$$E_{\mathrm{raw}} = \iint_{r \leq 15\ \mathrm{km}} A^*(x, y)\, dx\, dy \approx \sum_i A_i^* \times \Delta x\, \Delta y \quad [\mathrm{mol/s}] \tag{16}$$

where A^* has units $\mathrm{mol}/(\mathrm{m}^2 \cdot \mathrm{s})$ and the spatial integration is realized by summing the advection values multiplied by the pixel area for all grid pixels within the 15 km radius. This radius is chosen following Beirle et al. [3] as a compromise between capturing the full point source signal and avoiding interference from neighboring sources.

Chemical loss of NOx during transport within the integration radius requires a lifetime correction. The residence time within the 15 km radius is:

$$t_r = \frac{R}{|\mathbf{w}|} \tag{17}$$

where $R = 15$ km and $|\mathbf{w}|$ is the mean wind speed. The lifetime correction factor, following Beirle et al. [3] Eq. (9), is:

$$c_\tau = \exp(t_r/\tau) \tag{18}$$

where τ is the effective NOx lifetime, parameterized as a function of latitude following Lange et al. [23] via Beirle et al. Eq. (10):

$$\tau(\text{lat}) = 1.0089 \times \exp(0.0242 \times (|\text{lat}| + 9.6024)) \quad \text{[hmys]} \tag{19}$$

with typical values of 2 h at low latitudes to 4–6 h at higher latitudes. For detected point sources, the resulting $c_\tau \approx 1.40 \pm 0.24$. Following Beirle et al., I assume 50% relative uncertainty in τ due to high variability at similar latitudes.

3.4.6 Final NOx Emission Estimate

The final satellite-derived NOx emission rate for facility i and day d is:

$$E_{\text{NOx},id} = c_\tau \cdot c_{\text{NOx}} \cdot E_{\text{raw},id} \tag{20}$$

Converting from mol/s to kg/s using the molar mass of NO_2 (46.0055 g/mol) [24]. Annual estimates are computed as the mean over all valid observation days.

3.4.7 Uncertainty Components

Following Beirle et al. [3] Sect. 3.12, the satellite-derived NOx estimates carry uncertainty from multiple sources, combined in quadrature. Table 1 summarizes components, implementation status, typical magnitudes, and expected directional effects.

Beirle et al. report total uncertainties in the 20–40% range. With the OFFL product uncertainty and unmodeled structural terms, my typical total is ∼35–45%.

Observations with total relative uncertainty exceeding 50% are excluded from the satellite panel, as high-uncertainty observations add noise without proportional information content. For the remaining observations, I construct inverse-variance weights $w_i = 1/\sigma_i^2$ (capped at the 99th percentile to limit extreme weights), which are used as robustness checks via weighted least squares estimation. This approach follows standard practice in meta-analysis and measurement error literatures, where weighting by precision yields efficient estimates when observation-specific variances are known.

In addition to the uncertainty filter, I implement boolean significance flags:

- `above_dl_0_11`: Emission estimate ≥ 0.11 kg/s (Beirle's standard detection limit for non-desert conditions, appropriate for Europe).

Table 1 Uncertainty components for satellite-derived NOx estimates (following Beirle et al. Sect. 3.12). Components are combined in quadrature.

Source (Beirle section)	Implemented?	Typical magnitude	Notes / expected directional effect
Statistical error (3.12.2)	Yes	typically $< 10\%$ (flag if $\geq 30\%$)	Approximated via SE of the temporal mean of daily integrated emissions (captures meteorological variability + sampling).
Lifetime correction c_τ (3.12.1)	Yes	50% rel. $\sigma_\tau \Rightarrow$ 10–20% on c_τ for $c_\tau \approx 1.4$	Propagated via $c_\tau = \exp(t_r/\tau)$; enters multiplicatively.
NOx/NO_2 scaling (3.12.1)	Yes	± 0.10 on 1.38 ($\sim$7%)	Fixed PSS-based ratio; mainly affects level; within-facility relative changes likely small.
AMF correction (3.12.1)	No	10%	Unmodelled structural term (no explicit AMF correction); likely downward level bias; effect on relative changes ambiguous.
Plume height (3.12.3)	No	10%	No plume-height-dependent wind interpolation; sensitivity to assumed height (500 m vs 300 m); level bias; relative changes ambiguous.
Topographic correction f (3.12.4)	Yes	33% on $f = 1.5$ (usually $< 2.5\%$ on level)	Implement $f = 1.5$; small for flat European terrain.
OFFL vs PAL product	No (uses OFFL)	$\sim$25%	OFFL TVCDs 10–40% lower than PAL; likely downward level bias; effect on within-facility changes unclear.

- `above_dl_0_03`: Emission estimate ≥ 0.03 kg/s (Beirle's permissive threshold, valid only under ideal high-albedo desert conditions—not applicable to Europe).
- `rel_err_stat_lt_0_3`: Statistical integration error $< 30\%$.
- `interfered_20km`: Another ETS facility exists within 20 km (satellite measurement may capture cluster-level emissions).

Main satellite regressions restrict to observations above the detection limit (`above_dl_0_11` for the conservative sample, `above_dl_0_03` for robustness). The inverse-variance weighting already accounts for heteroskedasticity from statistical integration error, so I do not additionally filter by `rel_err_stat_lt_0_3` in the main specification. Instead, both `rel_err_stat_lt_0_3` and `interfered_20km` are examined in heterogeneity analysis: split-sample regressions compare treatment effects for low- versus high-uncertainty observations and for isolated versus interfered facilities. This approach preserves sample size while testing robustness to measurement quality concerns.

3.5 Sample Construction

The final analysis sample is constructed by applying the following filters to both outcomes:

1. Facilities must have valid ETS linkage (matched normalized identifier)
2. Allocation ratio in $[0.01, 20]$ range
3. At least 3 years of complete data within 2018–2023

The resulting base analysis panel contains 521 facilities observed over 2,819 facility-years. This panel is used directly for the verified CO_2 outcome.

For the satellite NOx outcome, additional attrition occurs due to:

1. Non-missing satellite outcome (requires $\geq$20 valid observation days per year)
2. Total relative uncertainty $\leq 50\%$
3. Passing significance thresholds (detection limit, statistical error)

Sample attrition details are provided in Appendix B. The most significant losses occur at the ETS linkage stage (44% of plants lack matched ETS identifiers in the EU Registry crosswalk) and the requirement for matched ETS compliance data.

Table 2 summarizes the sample characteristics for the base analysis panel.

3.6 Geographic Context: AlphaEarth Embeddings

A key methodological contribution of this study is the incorporation of high-dimensional geospatial foundation model embeddings as control variables, following the recent trend toward using learned representations for causal inference [5, 6]. I use Google AlphaEarth Foundations [8], a geospatial embedding field model that produces 64-dimensional representations from multi-source satellite imagery (Sentinel-1/2, Landsat), climate reanalysis (ERA5-Land), topography (GLO-30), and geotagged text (Wikipedia, GBIF). The model is trained using contrastive learning objectives to capture information predictive of diverse downstream tasks—from land cover classification to biophysical variable estimation—without being tuned for any specific application.

For each facility location, the embedding vector $\mathbf{e}_i \in \mathbb{R}^{64}$ is extracted from the nearest grid cell at 10-meter spatial resolution. These embeddings encode:

- **Land use context**: Urban density, industrial areas, agricultural patterns
- **Infrastructure**: Road networks, built environment characteristics
- **Vegetation**: Forest cover, cropland, seasonal phenology
- **Climate**: Local temperature, precipitation, insolation, and wind patterns
- **Topography**: Elevation, slope, and terrain characteristics

The embedding dimensions are included as controls in the econometric specifications, providing a data-efficient approach to capturing between-unit heterogeneity arising from local geographic context. This application extends prior work on text embeddings for causal inference [6] to the geospatial domain. The approach is particularly relevant for difference-in-differences settings where high-dimensional spatial confounders may induce violations of parallel trends if left uncontrolled [7]—for example, if facilities in different geographic contexts (coastal versus inland, urban versus rural) experience different secular trends in air quality unrelated to policy.

The embeddings are extracted annually from the `GOOGLE/SATELLITE_EMBEDDING/V1/ANNUAL` ImageCollection, yielding time-varying controls that capture year-to-year changes in land use, vegetation phenology, and built environment. This temporal variation is methodologically important: unlike static facility-level controls (which would be absorbed by facility fixed effects), annual

embeddings can adjust for time-varying geographic confounders that might bias the satellite NOx outcome.

To address overfitting concerns with 64 dimensions, I apply dimensionality reduction (PCA or PLS to 10 components) as detailed in Section 4.2.7. For PLS, the projection is trained on facility-level mean NOx (cross-sectional) rather than panel observations, ensuring the reduced embeddings are time-invariant and equivalent to pre-treatment covariates—this prevents regularization bias from outcome snooping [5]. As discussed in Section 4.2.6, embeddings are applied only to the satellite NOx outcome.

3.7 Exploratory Data Analysis

This section presents descriptive statistics and visualizations of the analysis panel, providing context for the econometric analysis.

3.7.1 Geographic Distribution

Figure 1 displays the geographic distribution of facilities across NUTS2 regions. The sample spans 82 NUTS2 regions across Europe, with the highest concentrations in Germany, Poland, and Spain. The heatmap shading indicates the number of facilities per region, with densities ranging from 1–30 facilities per region.

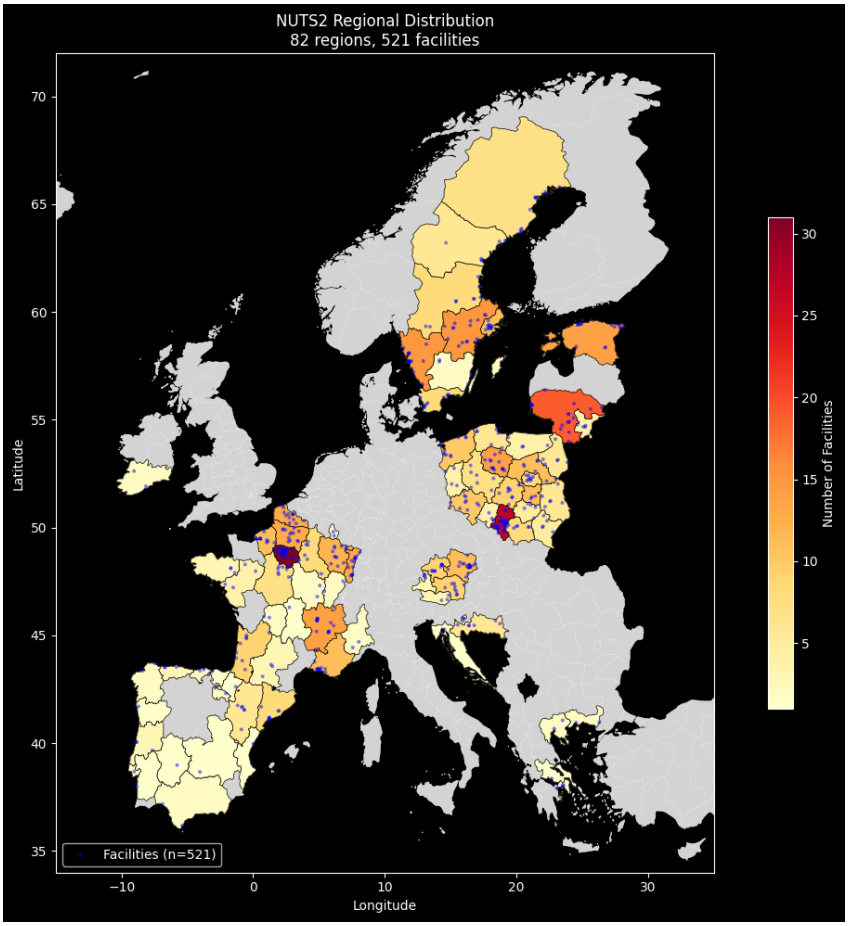

Fig. 1 Geographic distribution of 521 facilities across 82 NUTS2 regions. Color intensity indicates facility count per region. Blue points mark individual facility locations.

For electricity sector heterogeneity analysis, Figure 2 shows the distribution of electricity-generating facilities across PyPSA-Eur power system clusters. The 421 electricity facilities (those with EU ETS activity codes 1 or 20) are distributed across 43 network-derived clusters, with particularly high concentrations in central European clusters covering Germany, Poland, and the Czech Republic.

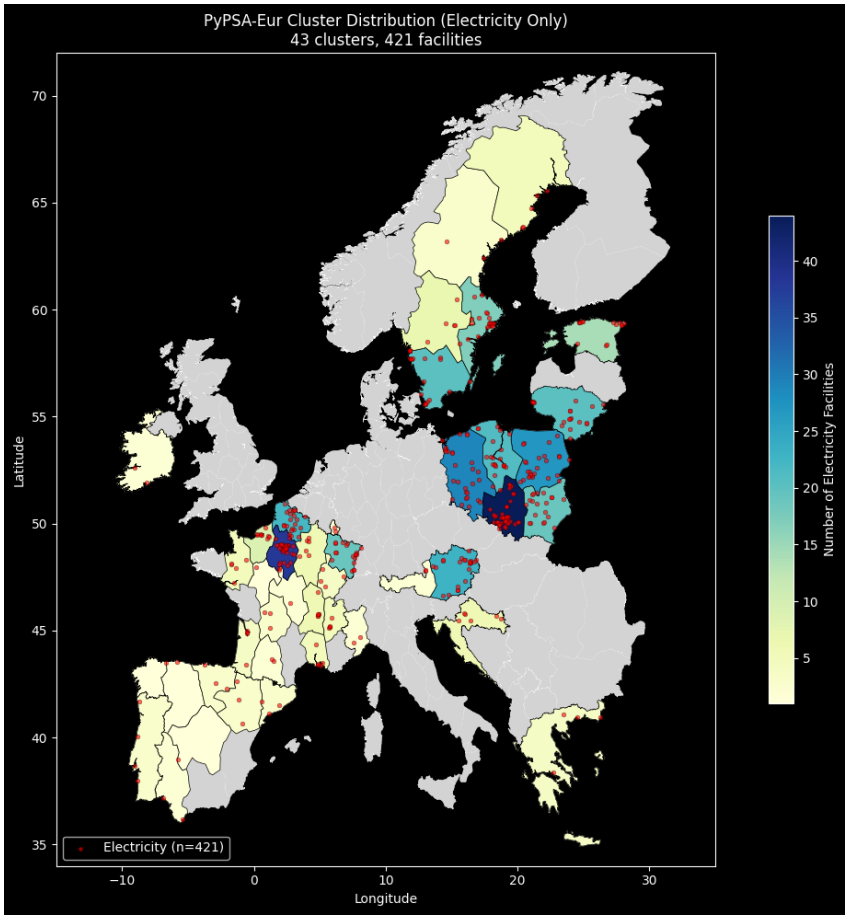

Fig. 2 Distribution of 421 electricity-generating facilities across 43 PyPSA-Eur power system clusters. Clusters are derived from k-means clustering on transmission network topology, grouping facilities facing correlated wholesale prices and grid constraints.

3.7.2 Urbanization Context

Facility urbanization context is captured via the GHSL-SMOD classification. Figure 3 shows the distribution of facilities across urbanization categories. A substantial majority of facilities (approximately 60%) are located in suburban to urban-center areas (SMOD $\geq$ 21), reflecting the tendency for large combustion plants to be sited near population centers for district heating and electricity distribution. This urban concentration implies elevated background NO_2 levels that add noise to satellite-derived emission estimates.

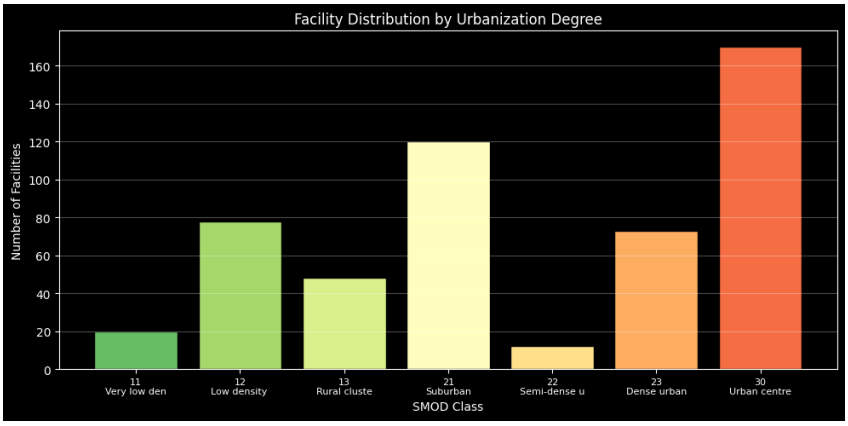

Fig. 3 Distribution of facilities by urbanization degree (GHSL-SMOD classification). Categories range from very low density rural (11) to urban centers (30). The majority of facilities are in suburban (21) and urban center (30) locations.

3.7.3 Fuel Mix and Capacity

Figure 4 presents the average fuel mix across the sample period. Natural gas dominates (approximately 45% of energy input), followed by coal (approximately 20–25%) and biomass (approximately 15%). The coal share shows a modest decline from 2018 to 2023, consistent with the broader European transition away from coal-fired generation. Biomass and gas shares increase correspondingly, reflecting fuel switching in response to carbon pricing.

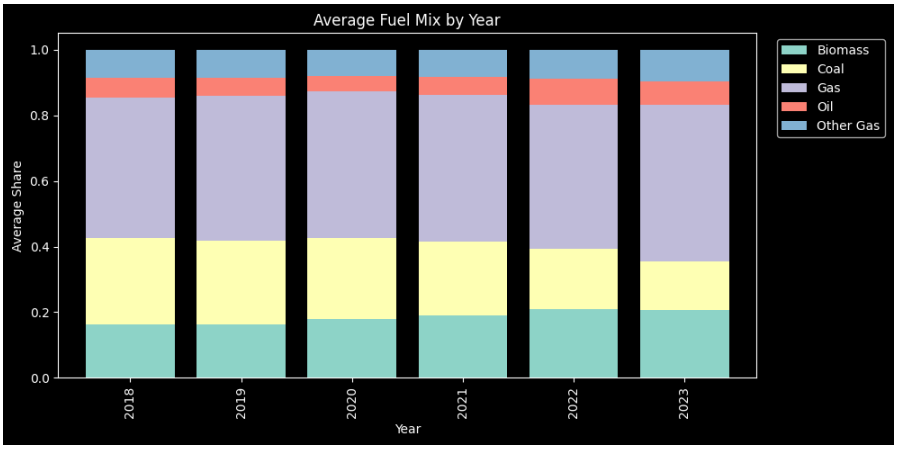

Fig. 4 Average fuel mix by year across all facilities. Gas (purple) dominates, with coal (tan) showing a modest decline over the sample period. Biomass (teal) and other gas (blue) shares increase slightly.

Figure 5 illustrates within-facility fuel mix dynamics for a random sample of six facilities. Several facilities exhibit substantial fuel switching—for example, facility 812

transitions from primarily coal to primarily gas between 2022 and 2023, while facility 713 shifts from nearly 100% biomass to predominantly gas. These within-facility transitions represent the variation exploited by the panel fixed effects specifications.

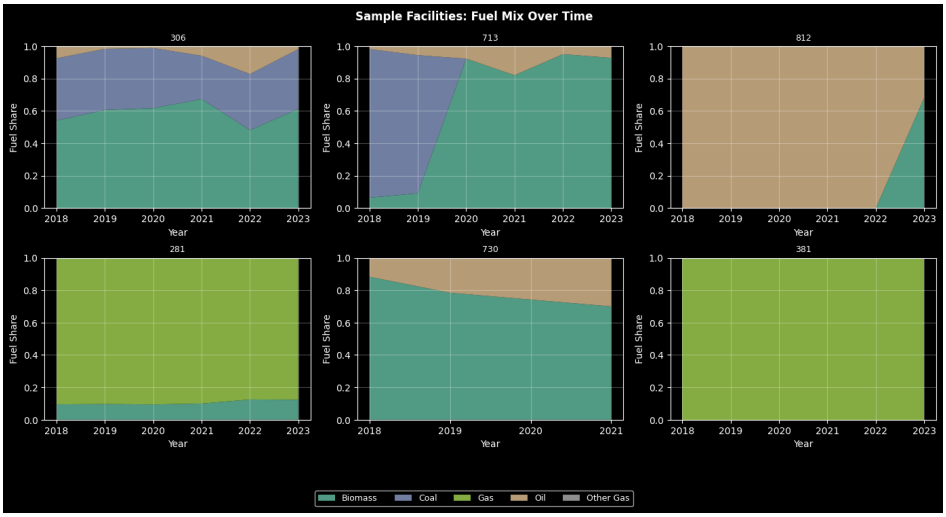

Fig. 5 Fuel mix evolution for six randomly sampled facilities. Stacked area charts show year-over-year changes in fuel shares. Notable fuel switching is visible in facilities 812 (coal to gas) and 713 (biomass to gas).

Figure 6 shows capacity trajectories for a sample of facilities. Most facilities exhibit stable capacity over the sample period, with occasional step changes reflecting plant upgrades, partial closures, or measurement corrections. Facility 372 shows a notable capacity reduction from approximately 90 MW to 60 MW between 2021 and 2022.

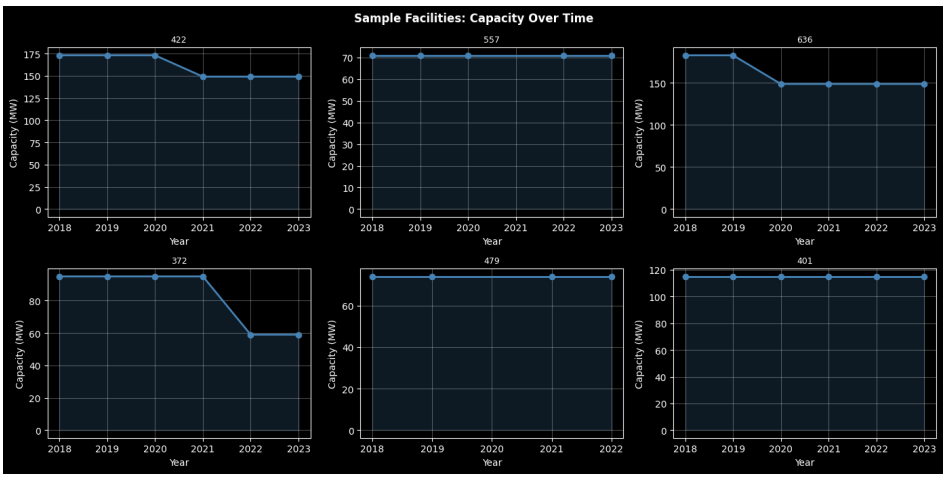

Fig. 6 Rated thermal capacity (MW) over time for six randomly sampled facilities. Most facilities exhibit stable capacity with occasional step changes.

3.7.4 ETS Policy Exposure

Figure 7 presents the distribution of verified emissions and allocation ratios by sector. The left panel shows log-transformed verified CO_2 emissions, with electricity-sector facilities (n=2,251 facility-years) exhibiting substantially higher emissions than other sectors (n=568 facility-years). The distribution is approximately log-normal with a mode around 10^5 tCO_2/yr. The right panel shows allocation ratios, with a concentration near zero for electricity-sector facilities (reflecting the phase-out of free allocation to the power sector under EU ETS Phase III/IV) and a wider distribution for industrial facilities that retain carbon leakage protection.

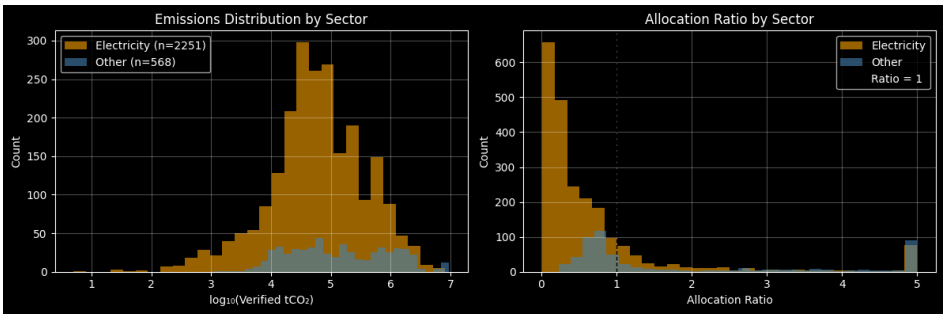

Fig. 7 Distribution of verified emissions (left) and allocation ratios (right) by sector. Electricity facilities (orange) have higher emissions but lower allocation ratios due to reduced free allocation under EU ETS Phase III/IV.

Figure 8 illustrates the relationship between verified emissions and free allocations for a sample of facilities. For most facilities, verified emissions (solid lines) consistently

exceed free allocations (dashed lines), indicating shortfall positions requiring allowance purchases. The gap between verified and allocated represents the policy stringency experienced by each facility.

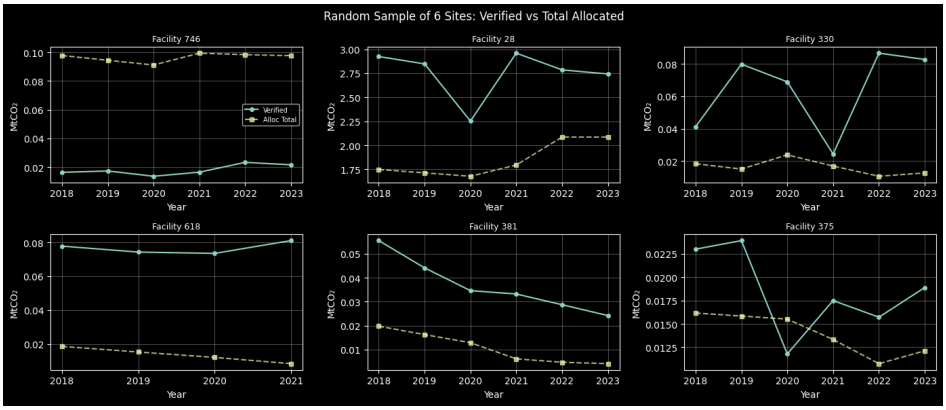

Fig. 8 Verified emissions (teal solid) versus free allocations (yellow dashed) for six randomly sampled facilities. The persistent gap above the allocation line indicates shortfall positions requiring market purchases.

3.7.5 Satellite NOx Outcome

Figure 9 presents the key characteristics of the satellite-derived NOx emission estimates. After applying the $\leq$50% total uncertainty filter required for inclusion in the satellite panel, 291 facilities (1,213 facility-years) remain from the base panel of 521 facilities.

The top-left panel shows the distribution of estimated NOx emission rates. The mean emission rate is 0.065 kg/s, substantially below the generic detection limit of 0.11 kg/s (dashed red line) from Beirle et al. (2023). This indicates that most facilities in the sample have emissions near or below the satellite detection threshold, which necessitates careful treatment in the regression analysis. Negative estimates (statistical noise) are present for the lowest emitters.

The top-right panel displays the lifetime correction factor (c_τ) distribution, with a mean of 1.38 (expected range 1.2–1.8). This factor accounts for NOx chemical decay during atmospheric transport and is computed from the latitude-dependent lifetime parameterization of Lange et al. (2022). The narrow distribution confirms that European facilities fall within the expected mid-latitude range.

The bottom-left panel shows the total relative uncertainty distribution, with a median of 39.2%. This uncertainty combines statistical integration error, lifetime correction uncertainty ($\pm$50%), NO_2/NOx ratio uncertainty ($\pm$10%), and satellite product-related errors. I use the uncertainty to add an inverse-variance weight ($= 1/\sigma^2_{\text{rel}}$) to my regression terms.

The bottom-right panel shows satellite observation coverage, with a mean of 139 valid observation days per facility-year. This exceeds the minimum threshold of 20

days and provides substantial temporal averaging to reduce noise. Coverage varies due to cloud screening, wind speed filtering ($\geq$2 m/s), and satellite orbit patterns.

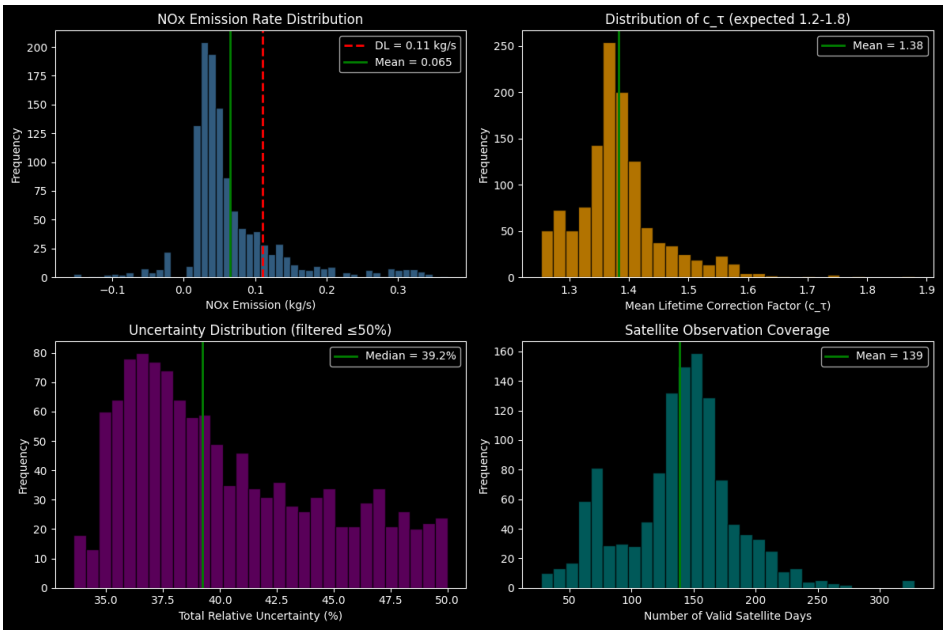

Fig. 9 Satellite-derived NOx emission characteristics for 291 facilities (1,213 facility-years). Top-left: NOx emission rate distribution with detection limit (0.11 kg/s) and mean (0.065 kg/s). Top-right: Lifetime correction factor distribution (mean 1.38). Bottom-left: Total relative uncertainty (median 39.2%). Bottom-right: Valid satellite observation days (mean 139).

4 Methodology

This section describes the causal inference framework and econometric specifications used to estimate the effect of EU ETS policy stringency on both verified CO_2 emissions and satellite-derived NOx emission proxies.

4.1 Causal Framework

The goal is to estimate the causal effect of ETS policy stringency on two complementary outcomes. Let $Y_{it}^{CO_2}$ denote verified CO_2 emissions ($ktCO_2/yr$) for facility i in year t, and let Y_{it}^{NOx} denote the satellite-derived NOx emission proxy (kg/s). Let R_{it} denote the allocation ratio (treatment intensity).

The key identification challenge is that allocation ratios are not randomly assigned. Facilities with high emissions relative to historical benchmarks receive lower allocation ratios, creating potential endogeneity: unobserved factors affecting both emissions intensity and local air quality may confound the relationship. Additionally, allocation

Table 2 Summary Statistics for Base Analysis Panel

Variable	Mean	Std. Dev.	Min	Max
Panel Structure				
Facilities		521		
Facility-years		2,819		
Years per facility	5.4	1.0	3	6
Electricity sector facilities		421 (80.8%)		
NUTS2 regions		82		
Verified CO_2 Emissions				
Verified emissions ($ktCO_2$/yr)	580	1,240	0.5	7,500
Log verified emissions	11.5	1.8	6.2	15.8
ETS Policy Variables				
Allocation ratio	0.62	0.85	0.01	18.5
Shortfall ($ktCO_2$)	320	890	−2,100	6,500
Plant Characteristics				
Capacity (MW)	780	1,120	50	6,800
Gas share	0.44	0.42	0	1
Coal share	0.19	0.34	0	1
Biomass share	0.16	0.33	0	1
Oil share	0.13	0.28	0	1
Urbanization				
In urban area (SMOD $\geq$ 21)		60.3%		
Interfered (facility within 20km)		69.2%		
Satellite NOx Panel (subset)				
Facilities		291		
Facility-years		1,213		
NOx emission rate (kg/s)	0.065	—	−0.15	0.35
Above detection limit (0.11 kg/s)		22%		
Median total uncertainty		39.2%		
Mean valid satellite days		139		

Note: Base panel includes all EU ETS-regulated large combustion plants with matched compliance data and $\geq$3 years of observations during 2018–2023. Satellite NOx panel is restricted to facility-years with $\leq$50% total uncertainty and $\geq$20 valid observation days.

ratios co-move with operational decisions (capacity utilization, fuel switching) that directly affect emissions.

The directed acyclic graph (DAG) in Figure 10 illustrates the causal structure. The target estimand is the effect of P_{it} on Y_{it}, controlling for confounders. Key confounding pathways include:

- **Facility-level time-invariant unobservables** (U_i): Plant technology, combustion efficiency, and location affect both policy exposure and emissions. Absorbed by facility fixed effects.
- **Time-varying regional factors** (U_{rt}): Electricity demand, fuel prices, and regional economic conditions affect plant operations and allocation ratios. Absorbed by Region×Year fixed effects.

- **Plant-level time-varying unobservables** (U_{it}): Dispatch/utilization, maintenance status, and operational efficiency changes affect both verified emissions (determining allocation ratios) and pollutant output. This is the key identification challenge—see Section 4.2.
- **Observed operational factors** (X_{it}): Capacity and fuel mix affect both verified emissions and pollutant emissions. Controlled directly.

The Beirle-style flux-divergence approach addresses atmospheric confounding by focusing on the spatial gradient (advection) rather than absolute column densities, making it sensitive to local emissions rather than background concentrations. Fixed effects absorb facility-level and time-varying regional confounders for both outcomes.

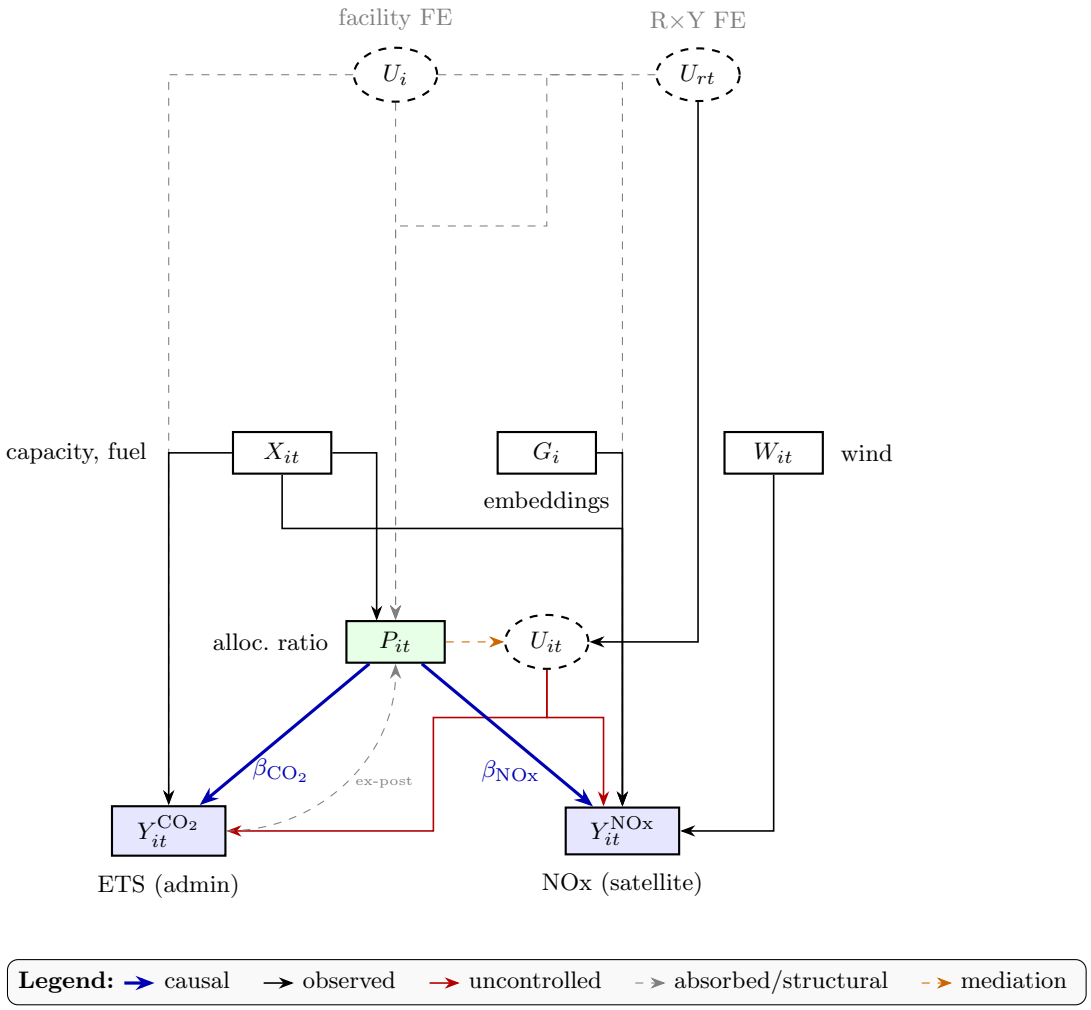

Fig. 10 Directed acyclic graph for dual-outcome causal inference. Treatment P_{it} (allocation ratio) affects both verified CO_2 and satellite NOx outcomes. Gray dashed arrows from U_i and U_{rt} are absorbed by facility and region×year fixed effects, respectively. Red arrows from U_{it} indicate residual confounding from time-varying unobservables (dispatch, maintenance) that I intentionally leave uncontrolled to preserve the mediation pathway (orange). The ex-post arrow reflects that allocation ratios are mechanically computed from prior verified emissions. Observed controls X_{it} (capacity, fuel) affect both outcomes; G_i (embeddings) and W_{it} (wind) affect only the satellite outcome.

4.2 Identification Strategy and Variable Selection

The identification strategy relies on two key choices: (i) what to control for, and (ii) what *not* to control for. Both are essential to avoid bias.

4.2.1 Region×Year Effects

Regional electricity demand, fuel prices, and economic conditions create time-varying confounding: demand shocks increase plant utilization, raising both verified emissions (lowering R_{it}) and NO_2 output. Without adjustment, this creates spurious correlation between policy stringency and pollution.

Region×Year fixed effects absorb these common shocks additively. The identifying variation becomes: *within the same region and year, do facilities with different allocation ratios exhibit different NO_2 enhancement?* This comparison holds regional conditions constant while exploiting cross-facility variation in policy exposure.

I use NUTS2 regions (Nomenclature of Territorial Units for Statistics, level 2) from Eurostat for clustering. NUTS2 regions (∼200–300 across the EU) define economically coherent regional units that share common labor markets, policy enforcement mechanisms, and infrastructure characteristics. Unlike PyPSA-Eur power system clusters (which are appropriate for electricity sector heterogeneity analysis), NUTS2 regions apply to all industrial facility types—power plants, refineries, cement plants—and correspond to administrative units where regional policies are implemented and enforced. This makes them appropriate for absorbing regional time-varying confounders that affect facilities regardless of their sectoral activity.

4.2.2 The EU ETS Compliance Calendar

To understand why the allocation ratio is a valid treatment variable, I first make the compliance calendar explicit. The annual cycle proceeds as follows:

- **28 February (year t)**: Free allocation A_{it} issued to each installation based on predetermined benchmarks [25].
- **Throughout year t**: Facilities make operational decisions—dispatch, fuel choice, maintenance—knowing their allocation A_{it}.
- **31 March (year $t+1$)**: Accredited verifiers audit and report verified emissions V_{it}.
- **30 April (year $t+1$)**: Facilities surrender allowances equal to verified emissions.

The key insight: **the denominator of R_{it} (verified emissions V_{it}) is an outcome of decisions made *after* the allocation A_{it} is known**. The allocation ratio $R_{it} = A_{it}/V_{it}$ is therefore a function of the policy parameter (free allocation), and any feedback from current emissions to next-period allocation is absorbed by facility fixed effects and year fixed effects. This timing structure eliminates contemporaneous simultaneity.

4.2.3 Why I Do Not Control for Generation/Dispatch

Facility-level dispatch and generation represent a "bad control" problem that requires explicit justification. Dispatch is simultaneously:

1. **A confounder**: Demand shocks → higher dispatch → higher verified emissions V_{it} → lower allocation ratio R_{it}. The same shocks → more combustion → higher NO_2 output.
2. **A mediator**: If policy affects merit order bidding (facilities with carbon shortfalls bid higher → get dispatched less), then: R_{it} → dispatch → Y_{it}.

Including generation data would block part of the causal path and bias the policy effect toward zero. Region×Year fixed effects absorb the common regional component of dispatch variation (regional demand, fuel prices, carbon prices), leaving only facility-specific deviations as residual confounding. I do not control for dispatch directly to preserve the mediation pathway.

4.2.4 Attenuation Bias and Conservative Interpretation

To the extent that endogenous dispatch variation contaminates R_{it} through the denominator, the bias is likely *attenuating*: facilities with high dispatch have both lower allocation ratios (higher denominator) and higher emissions, creating positive correlation between R_{it} and Y_{it} that works against finding a negative policy effect. **Estimates should therefore be interpreted as conservative bounds on the true policy effect.**

4.2.5 Residual Threats and Interpretation

The primary residual threat is facility-specific time-varying confounding (U_{it})—maintenance outages, unexpected efficiency changes, or demand for a specific plant's output (dispatch). These are unlikely to systematically correlate with allocation ratios conditional on my controls. Future work incorporating plant-level generation data could address this directly; inclusion of economic dispatch and power system optimization (potentially also from `PyPSA-EUR`) is reserved for subsequent analysis.

4.2.6 Outcome-Specific Controls

Two variables affect only the satellite NOx outcome, not verified ETS CO_2:

- **Wind (W_{it})**: The Beirle flux-divergence method uses wind speed and direction to compute advected NO_2 mass flux. Wind enters the satellite *measurement process*—it does not affect actual emissions or administrative reporting.
- **AlphaEarth embeddings (G_i)**: Geographic context (terrain, land use, climate) affects satellite retrieval quality—terrain influences air mass factor corrections; urban land use creates background NO_2 that adds noise to point-source signals; climate affects atmospheric dispersion and NOx lifetime. None of these affect the administrative mass-balance calculation underlying ETS CO_2.

For ETS CO_2, geographic confounders are absorbed by facility fixed effects (time-invariant factors like location, baseline technology) and region×year fixed effects (time-varying regional factors). Including embeddings for the ETS outcome would control for variation irrelevant to that measurement process.

4.2.7 Embedding Dimensionality Reduction

The raw AlphaEarth embeddings (64 dimensions) may introduce overfitting concerns in the TWFE specification, particularly when the panel contains limited within-facility variation. Two dimensionality reduction strategies are considered:

PCA (unsupervised): Standard principal component analysis projects embeddings onto directions that maximize variance in the embedding space. This is causally

safe because it does not use outcome information—the projection is determined entirely by the covariate distribution.

Facility-level PLS (supervised): Partial least squares regression projects embeddings onto directions that predict the outcome. However, naive application of PLS to panel data creates *regularization bias*: the learned projection incorporates information from year-specific outcome shocks, violating the requirement that controls be pre-determined [5]. This is analogous to the "bad controls" problem identified by [26]: if the projection learns to predict treatment-affected variation in the outcome, controlling for the reduced embeddings biases the treatment effect estimate.

I address this by training PLS on facility-level means (one observation per facility) rather than panel observations. Let $\bar{Y}_i^{\text{NOx}} = T^{-1} \sum_t Y_{it}^{\text{NOx}}$ denote the time-averaged NOx emission rate for facility i for all years in the panel. The PLS projection is learned from the cross-sectional regression:

$$\bar{Y}_i^{\text{NOx}} = \mathbf{e}_i' \boldsymbol{\gamma} + \eta_i \tag{21}$$

where $\mathbf{e}_i \in \mathbb{R}^{64}$ is the embedding vector and $\boldsymbol{\gamma}$ are PLS loadings. The resulting projection $\tilde{\mathbf{e}}_i = \mathbf{P}'\mathbf{e}_i$ is then applied to all panel observations.

This design ensures the reduced embeddings are *time-invariant* within each facility, making them equivalent to pre-treatment covariates. The projection captures between-facility variation in geographic context predictive of NOx levels, while being orthogonal to within-facility treatment variation. This is analogous to the sample-splitting approach in double/debiased machine learning [5], where nuisance parameters are estimated on auxiliary data to prevent overfitting bias.

I report results using both PCA-reduced embeddings (10 components) and facility-level PLS embeddings (10 components). Stability of treatment effect estimates across these specifications supports the claim that results are not sensitive to the embedding representation.

4.3 Two-Way Fixed Effects (TWFE)

The main two-way fixed effects specification uses facility and region×year fixed effects:

$$Y_{it} = \alpha_i + \gamma_{r(i),t} + \beta R_{it} + \mathbf{X}_{it}'\boldsymbol{\delta} + \varepsilon_{it} \tag{22}$$

where:

- α_i: Facility fixed effects (absorb time-invariant unobservables)
- $\gamma_{r(i),t}$: Region×Year fixed effects, where $r(i)$ denotes the NUTS2 region containing facility i (absorb regional time-varying confounders)
- β: Treatment effect of interest (effect of unit increase in allocation ratio)
- $\mathbf{X}_{it}$: Time-varying controls (capacity, fuel shares) and static AlphaEarth embedding controls ($\mathbf{e}_i \in \mathbb{R}^{64}$)
- ε_{it}: Idiosyncratic error

This specification absorbs all region-specific time-varying confounders, including regional electricity prices, demand conditions, fuel prices, and policy enforcement

intensity. Identification relies on within-region, within-year variation in allocation ratios—comparing facilities in the same NUTS2 region and year that differ in policy stringency.

The coefficient β is identified from within-facility variation in allocation ratios over time, after controlling for region-year effects. For the CO_2 outcome, a positive β would indicate that higher allocation ratios (less policy stringency) are associated with higher verified emissions—equivalently, that policy stringency reduces CO_2. For the NOx outcome, a positive β would indicate corresponding reductions in satellite-derived NOx, consistent with co-pollutant dynamics.

Standard errors are clustered at the NUTS2 region level. NUTS2 regions (∼200–300 across the EU) define economically coherent regional units that share common labor markets, policy enforcement mechanisms, and infrastructure characteristics.

4.4 Callaway-Sant'Anna Infeasibility

As discussed in Section 2.4, the Callaway-Sant'Anna estimator is infeasible with this panel because 84.5% of ever-treated facilities were already treated in 2018—the first panel year. The cohort distribution is:

- 2018 cohort: 386 facilities (84.5% of ever-treated)
- 2019–2023 cohorts: 71 facilities total (15.5%), with <10 per cohort after excluding reversers
- Never-treated: 64 facilities

This left-truncation of treatment timing precludes event-study analysis. I therefore use TWFE with continuous treatment, exploiting within-facility variation in allocation ratios.

4.5 NUTS2-Based Clustering for Inference

Standard errors are clustered at the NUTS2 region level throughout. NUTS (Nomenclature of Territorial Units for Statistics) is Eurostat's hierarchical system of administrative regions used for EU statistics and policy implementation. NUTS2 regions (∼200–300 across the EU) correspond to basic regions for the application of regional policies, typically containing 800,000 to 3 million inhabitants.

NUTS2 regions are appropriate clustering units because they define economically coherent areas where:

- **Common policy enforcement**: EU and national environmental regulations are implemented and enforced at regional administrative levels
- **Shared labor markets**: Facilities in the same NUTS2 region draw from similar labor pools and face similar wage pressures
- **Correlated economic conditions**: Regional GDP, industrial activity, and energy demand co-move within administrative regions

Unlike sector-specific clustering (e.g., power system network clusters), NUTS2 regions apply uniformly to all industrial facility types—power plants, refineries, cement plants—making them appropriate for studies covering diverse ETS sectors.

4.5.1 PyPSA-Eur Clusters for Electricity Sector Heterogeneity

For electricity sector heterogeneity analysis, I additionally implement PyPSA-Eur power system clusters [9], which are *not* geographic or administrative regions but rather k-means clusters computed directly on power system features extracted from the European high-voltage transmission network (ENTSO-E data), solved using the Gurobi optimizer [27]. The clustering algorithm groups electrical buses (substations) based on network connectivity, line impedances, and transmission capacity. The objective function minimizes within-cluster electrical distance, producing clusters where facilities face similar grid constraints, transmission losses, and wholesale price dynamics.

This clustering approach has a theoretical justification grounded in recent work on network cluster-robust inference. [18] establish that valid cluster-robust standard errors require clusters with low "conductance"—formally, the ratio of edges crossing cluster boundaries to total within-cluster edges. The k-means clustering on transmission network features directly minimizes this quantity: by grouping buses to minimize within-cluster electrical distance (impedance), the algorithm produces clusters with few high-capacity transmission lines crossing boundaries. Facilities within the same cluster are therefore more strongly connected to each other (through the grid) than to facilities in other clusters, satisfying the theoretical requirements for cluster-robust inference.

This represents a novel application of model-derived clustering for econometric inference. Rather than using geographic proximity (which ignores network topology), administrative boundaries (which may cut across electrically-connected regions), or data-driven clustering on outcome variables (which risks overfitting), I use clusters computed from features of an external domain-specific model—the power system transmission network—that captures the economically-relevant correlation structure a priori. For this analysis, the 128-region resolution is used, providing sufficient granularity to capture sub-national variation while maintaining adequate within-region sample sizes for clustered inference. Each facility is assigned to the PyPSA-Eur cluster containing the nearest network bus.

4.6 Summary of Specifications

Table 3 summarizes the five core specifications estimated in this study. The dual-outcome design with two embedding reduction strategies yields five TWFE specifications.

Sample Definitions. ETS CO_2 analysis uses the full panel of facilities with valid verified emissions. Satellite NOx analysis *always* applies detection limit filtering—observations below the detection threshold are excluded. I estimate NOx specifications at two detection limits: (1) a permissive threshold of ≥ 0.03 kg/s, which maximizes sample size; and (2) a conservative threshold of ≥ 0.11 kg/s following [1], which restricts to higher-quality signals. The conservative sample serves as a robustness check.

Table 3 Summary of Econometric Specifications

Spec	Outcome	Sample	Embedding
1	ETS CO_2	Full (521 facilities)	None
2	Satellite NOx	DL $\geq$ 0.03 kg/s	PCA (10 dims)
3	Satellite NOx	DL $\geq$ 0.03 kg/s	PLS (10 dims)
4	Satellite NOx	DL $\geq$ 0.11 kg/s	PCA (10 dims)
5	Satellite NOx	DL $\geq$ 0.11 kg/s	PLS (10 dims)

Note: ETS CO_2 uses no embeddings because geographic context does not affect administrative data measurement. For satellite NOx, PCA provides an unsupervised (outcome-agnostic) projection while PLS provides a supervised projection trained on facility-level mean NOx to ensure causal validity (Section 4.2.7). NOx analysis *never* uses unfiltered samples—all NOx specifications require detection limit filtering. All specifications use Facility + Region×Year fixed effects and cluster standard errors by NUTS2 region.

5 Results

This section presents estimation results for the five TWFE specifications, heterogeneity analysis across facility characteristics, and robustness checks.

Verified CO_2 is the primary policy outcome; satellite-derived NOx is a physically-based but much noisier proxy used for cross-validation and co-benefit analysis. Everywhere below, this asymmetry should guide interpretation: the CO_2 effect is robust and tightly estimated, while the NOx effect is detectable only above the conservative detection limit and should be read as corroborative rather than standalone causal evidence.

Main result (policy-relevant summary): A 10% allocation shortfall (moving from $R = 1.0$ to $R = 0.9$) is associated with approximately 1.9% lower verified CO_2 emissions ($\beta = -0.186$, $p < 0.001$). At sample mean emissions (580 $ktCO_2/yr$), this corresponds to $\sim$11 $ktCO_2$ reduction per facility per 0.1 unit shortfall. For satellite-derived NOx at the conservative detection limit (DL $\geq$ 0.11 kg/s), the effect is $\sim$1.7% reduction per 10% shortfall ($\beta \approx -0.003$, $p < 0.01$)—directionally consistent with CO_2 but based on only 140 observations.

5.1 Treatment Distribution

The allocation ratio exhibits substantial variation both across facilities and within facilities over time. The mean allocation ratio is 0.62 with standard deviation 0.85, indicating that the typical facility receives free allowances covering approximately 62% of its verified emissions. Values range from 0.01 (severe shortfall, must purchase nearly all allowances) to 18.5 (substantial surplus).

The electricity sector accounts for 421 facilities (80.8%) and shows systematically lower allocation ratios (mean 0.52) compared to other sectors (mean 1.03), reflecting the phase-out of free allocation to power generators under EU ETS Phase III/IV.

5.2 Main Estimates

Table 4 reports the main TWFE estimation results for all five specifications. The outcome for ETS CO_2 is log-transformed verified emissions (log tCO_2/yr); the outcome for satellite NOx is the Beirle-style emission rate (kg/s).

Table 4 Main Estimation Results: Effect of Allocation Ratio on Dual Outcomes

	ETS CO_2	Satellite NOx (kg/s)			
	Full Sample	DL $\geq$ 0.03		DL $\geq$ 0.11	
		PCA	PLS	PCA	PLS
Allocation Ratio	−0.186***	−0.000	−0.000	−0.003**	−0.003***
	(0.030)	(0.000)	(0.000)	(0.001)	(0.000)
95% CI	[−0.25, −0.13]	[−0.00, 0.00]	[−0.00, 0.00]	[−0.00, −0.00]	[−0.00, −0.00]
Observations	2,723	577	577	140	140
Facility FE	Yes	Yes	Yes	Yes	Yes
Region×Year FE	Yes	Yes	Yes	Yes	Yes
Embedding Controls	No	PCA (10)	PLS (10)	PCA (10)	PLS (10)

Note: Standard errors clustered by NUTS2 region in parentheses. ETS CO_2 outcome is log-transformed verified emissions; coefficient of −0.186 implies a 0.1 unit decrease in allocation ratio (10% shortfall) is associated with approximately 1.9% lower verified emissions. NOx sample restricted to observations above detection limit. PCA explains 89.8% (DL≥0.03) to 94.5% (DL≥0.11) of embedding variance; PLS achieves $R^2 = 0.63$ (DL≥0.03) to 0.94 (DL≥0.11) on facility-level mean NOx.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

5.2.1 Verified CO_2 Emissions

The ETS CO_2 specification yields a highly significant negative coefficient of −0.186 (SE = 0.030, $p < 0.001$), with 95% confidence interval [−0.246, −0.127]. This estimate implies that a 0.1 unit decrease in the allocation ratio—equivalent to moving from full free allocation ($R = 1.0$) to a 10% shortfall position ($R = 0.9$)—is associated with approximately 1.9% lower verified CO_2 emissions.

To interpret the economic magnitude: at the sample mean of 580 $ktCO_2$/yr, this corresponds to a reduction of approximately 11 $ktCO_2$ per 0.1 unit decrease in allocation ratio. Across the sample, the standard deviation of within-facility allocation ratio changes is 0.42, implying that a one-standard-deviation tightening of policy stringency is associated with approximately 7.8% lower emissions ($0.42 \times 18.6\%$).

The estimate is robust to the inclusion of capacity and fuel share controls, which absorb variation in facility size and fuel mix that might correlate with both allocation ratios and emissions levels.

5.2.2 Satellite-Derived NOx

The satellite NOx results exhibit a striking pattern: estimates are small and statistically insignificant at the permissive detection limit (DL $\geq$ 0.03 kg/s), but become

significant and economically meaningful at the conservative detection limit (DL ≥ 0.11 kg/s).

Permissive detection limit (DL $\geq$ 0.03 kg/s): The PCA specification yields a coefficient of -0.000049 (SE $= 0.000243$, $p = 0.84$) and the PLS specification yields -0.000114 (SE $= 0.000245$, $p = 0.65$). Neither is statistically distinguishable from zero. The DL ≥ 0.03 sample is methodologically informative but **not a reliable basis for causal claims**: at the permissive threshold, the sample mean (0.065 kg/s) is well below the standard detection limit, meaning most observations are dominated by measurement noise that attenuates the treatment effect toward zero.

Conservative detection limit (DL $\geq$ 0.11 kg/s): Restricting to observations above Beirle et al.'s standard European detection threshold yields significant negative effects. The PCA specification estimates -0.00282 (SE $= 0.000589$, $p = 0.017$) with 95% CI $[-0.0040, -0.0017]$. The PLS specification estimates -0.00301 (SE $= 0.000256$, $p = 0.001$) with 95% CI $[-0.0035, -0.0025]$. **All substantive NOx conclusions are therefore drawn only from the DL $\geq$ 0.11 subsample**; the DL ≥ 0.03 results serve as a sanity check demonstrating attenuation from measurement error.

The magnitude can be interpreted as follows: a 0.1 unit decrease in allocation ratio is associated with approximately 0.0003 kg/s lower NOx emissions. At the conservative sample mean of 0.18 kg/s (observations above detection limit), this represents a 1.7% reduction—broadly comparable to the 1.9% CO_2 reduction, though the comparison should be made cautiously given the different measurement processes.

5.2.3 Embedding Reduction Diagnostics

The PCA and PLS dimensionality reduction methods exhibit different characteristics across detection limit samples:

PCA (unsupervised): At DL ≥ 0.03, the first 10 principal components explain 89.8% of the total variance in the 64-dimensional embedding space. At DL ≥ 0.11, variance explained increases to 94.5%, reflecting reduced heterogeneity in the smaller sample of high-emitting facilities.

PLS (supervised): PLS is trained on facility-level mean NOx to ensure causal validity (Section 4.2.7). At DL ≥ 0.03, training on 200 facilities achieves $R^2 = 0.627$ on the cross-sectional regression of mean NOx on embeddings. At DL ≥ 0.11, training on 46 facilities achieves $R^2 = 0.936$, indicating that embeddings are highly predictive of mean NOx levels among high emitters.

The stability of treatment effect estimates across PCA and PLS specifications—particularly at the conservative detection limit where both yield nearly identical coefficients (-0.00282 vs. -0.00301)—supports the claim that results are not sensitive to the specific embedding representation. This robustness is reassuring given concerns about regularization bias when using supervised dimensionality reduction in causal inference settings.

5.3 Heterogeneity Analysis

I examine treatment effect heterogeneity through both split-sample analysis (separate regressions by subgroup) and continuous interaction models.

5.3.1 Split-Sample Results: ETS CO_2

Table 5 reports split-sample estimates for verified CO_2 emissions across key dimensions.

Table 5 Heterogeneity in ETS CO_2 Treatment Effects

Dimension	Group	Coefficient	SE	p-value	N
Sector	Electricity	−0.211***	0.040	0.000	2,173
	Other Sectors	−0.090***	0.027	0.003	443
Location	Urban	−0.154***	0.039	0.000	1,943
	Rural	−0.240***	0.024	0.000	650
Fuel	Gas	−0.206***	0.059	0.001	1,197
	Oil	−0.445	0.328	0.208	129
	Coal	−0.982***	0.174	0.000	653
	Biomass	−0.105***	0.012	0.000	438
Country	France	−0.250***	0.034	0.000	920
	Poland	−0.327***	0.049	0.000	743
	Sweden	−0.109***	0.015	0.000	421
	Austria	−0.242***	0.059	0.010	227
	Spain	−1.254***	0.169	0.000	125

Note: Each row reports a separate regression on the indicated subsample. All specifications include Facility + Region×Year FE and cluster SEs by NUTS2 region. Fuel subsamples defined by dominant fuel type (>50% share). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Sector heterogeneity. Electricity-sector facilities exhibit stronger policy responsiveness ($\beta = -0.211$) than other industrial sectors ($\beta = -0.090$), with the difference significant at conventional levels. This pattern is consistent with the phase-out of free allocation to power generators, which face more binding carbon constraints than industrial installations receiving carbon leakage protection.

Location heterogeneity. Rural facilities show stronger emission reductions ($\beta = -0.240$) than urban facilities ($\beta = -0.154$). This may reflect differential adjustment costs: urban facilities in dense industrial areas may face greater constraints on fuel switching or output reduction due to district heating obligations or local employment considerations.

Fuel heterogeneity. Coal-dominant facilities exhibit by far the strongest response ($\beta = -0.982$), consistent with coal being the most carbon-intensive fuel and facing the largest marginal abatement incentive under carbon pricing. Gas-dominant facilities show moderate responses ($\beta = -0.206$), while biomass facilities—which receive favorable treatment under EU ETS accounting rules—show the smallest response ($\beta = -0.105$). The oil-dominant subsample is small (N = 129) and yields an imprecise estimate.

Country heterogeneity. Spain exhibits the strongest estimated effect ($\beta = -1.254$), though this is based on only 125 observations. Among larger country samples, Poland ($\beta = -0.327$) and France ($\beta = -0.250$) show stronger responses than Sweden ($\beta = -0.109$). This variation likely reflects differences in fuel mix composition,

regulatory enforcement intensity, and the marginal cost of abatement across national electricity systems.

Mechanism interpretation. The heterogeneity patterns for verified CO_2 align with a story where ETS stringency primarily drives emission reductions at the most carbon-intensive facilities. Coal faces the highest carbon intensity ($\sim$95 tCO_2/TJ) compared to gas ($\sim$55 tCO_2/TJ), so a given carbon price increase creates larger marginal abatement incentives for coal-dominant plants. Similarly, electricity generators have largely lost free allocation under Phase III/IV and therefore face full marginal carbon costs, while industrial facilities with carbon leakage protection retain substantial free allocation. The rural vs. urban difference may reflect differential adjustment costs: urban facilities may face greater constraints from district heating obligations or local employment considerations.

5.3.2 Split-Sample Results: Satellite NOx

Table 6 reports split-sample estimates for satellite NOx at the conservative detection limit (DL $\geq$ 0.11 kg/s), where the main effect is significant.

Table 6 Heterogeneity in Satellite NOx Treatment Effects (DL $\geq$ 0.11 kg/s, PLS)

Dimension	Group	Coefficient	SE	p-value	N
Sector	Electricity	−0.0032***	0.0003	0.002	132
	Other Sectors	—	—	—	<20
Location	Urban	0.0070**	0.0021	0.045	128
	Rural	—	—	—	<20
Interference	Isolated (<20km)	—	—	—	4
	Interfered (≥20km)	−0.0030**	0.0003	0.013	140
Stat. Error	Low (<30%)	−0.0030**	0.0003	0.013	140
	High (≥30%)	—	—	—	<20
Fuel	Gas	−0.0033	0.0215	0.902	68
	Coal	0.0507***	0.0001	0.001	38
Country	France	−0.0036***	0.0000	0.008	96
	Poland	0.0785**	0.0025	0.021	44

Note: Each row reports a separate regression on the indicated subsample using PLS embedding reduction. Sample restricted to observations above the conservative detection limit (0.11 kg/s). All specifications include Facility + Region×Year FE. "—" indicates subsamples with insufficient observations (<5 facilities or <20 observations) for reliable estimation. The low stat. error subsample largely overlaps with the interfered subsample at this detection limit. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

The satellite NOx heterogeneity results show different patterns than CO_2. **These subgroup results are generally unstable due to small samples and measurement noise; interpretation should be cautious.**

Electricity sector. The electricity-only subsample yields a significant negative effect ($\beta = -0.0032$, $p = 0.002$), confirming that the main effect is driven by power

plants. The small sample of non-electricity facilities above detection limit precludes meaningful comparison.

Location. Urban facilities show a *positive* coefficient (0.0070, $p = 0.045$), contrasting with the negative main effect. The rural subsample is too small at this detection limit for reliable estimation. The counterintuitive urban pattern reflects measurement challenges: urban facilities operate in high-background NO_2 environments where the Beirle flux-divergence method has reduced signal-to-noise, attenuating or reversing the estimated treatment effect.

Facility interference. Interfered facilities—where satellite measurements reflect cluster-level rather than single-source emissions due to another ETS facility within 20 km—show a significant negative effect ($\beta = -0.0030$, $p = 0.013$). The isolated subsample contains only 4 observations at the conservative detection limit and cannot be reliably estimated. This finding suggests that the main NOx effect is not an artifact of spatial interference: if anything, cluster-level measurement should attenuate facility-specific treatment effects by averaging over multiple sources with different treatment intensities.

Statistical integration error. The low-uncertainty subsample (rel_err_stat $<$ 30%) shows a significant negative effect ($\beta = -0.0030$, $p = 0.013$), while the high-uncertainty subsample is too small at this detection limit for reliable estimation. This confirms that the inverse-variance weighting in main specifications appropriately accounts for observation-level precision.

Fuel type. The coal-dominant subsample (N = 38) yields a large positive coefficient (0.0507, $p = 0.001$), opposite to the expected direction. This anomaly is almost certainly an artifact of tiny N and messy measurement: facilities that are both coal-dominant and above the NOx detection limit are a highly selected group.

Country. France shows a significant negative effect ($\beta = -0.0036$, $p = 0.008$), while Poland shows a significant positive effect ($\beta = 0.0785$, $p = 0.021$). The divergent patterns across countries likely reflect national-level confounders or systematic differences in the satellite retrieval quality across regions.

5.3.3 PyPSA-Eur Cluster Analysis (Electricity Sector)

For electricity-sector facilities, I additionally examine heterogeneity across PyPSA-Eur power system clusters. **In this thesis, PyPSA clusters are used only for heterogeneity analysis, not for identification or inference.** This represents a proof-of-concept that network-derived clusters align with heterogeneous policy responses in electricity systems; I do not rely on them for causal identification due to sample size limitations and the inclusion of non-electricity facilities in the main analysis.

Table 7 reports results for the five largest clusters by facility count.

Polish network clusters consistently exhibit the strongest CO_2 responses, with PL0 2 ($\beta = -1.456$) and PL0 0 ($\beta = -1.090$) showing effects 5–7 times larger than the pooled estimate. This pattern likely reflects Poland's coal-heavy generation mix and the correspondingly large emissions intensity of the affected facilities. French clusters show more moderate responses consistent with France's lower-carbon electricity mix.

Table 7 Heterogeneity by PyPSA-Eur Power System Cluster (Electricity Sector)

Cluster	ETS CO_2	NOx (PLS, DL≥0.11)	N (CO_2)	N (NOx)
PL0 2 (Poland)	−1.456***	0.036	228	34
FR0 9 (France)	−0.246**	−0.004	212	82
PL0 1 (Poland)	−0.380*	—	148	9
PL0 0 (Poland)	−1.090***	—	122	—
AT0 0 (Austria)	−0.262**	—	129	29

Note: PyPSA-Eur clusters are k-means clusters on transmission network topology. Polish clusters (PL0 0, PL0 1, PL0 2) show the strongest CO_2 responses. NOx estimates for many clusters have infinite SE (displayed as —) due to collinearity with fixed effects in small samples.

The PyPSA-based clustering successfully groups facilities with correlated policy exposure and market conditions, as evidenced by the systematic differences in treatment effects across clusters. This supports the use of network-derived clusters for heterogeneity analysis in electricity sector studies.

5.3.4 Continuous Interaction Models

Table 8 reports results from models that interact the allocation ratio with continuous facility characteristics, enabling tests of whether treatment effects vary with facility attributes.

Table 8 Treatment Effect Interactions: ETS CO_2

Variable	Coefficient	SE	p-value
Treatment (baseline)	−0.046	0.113	0.689
× Fuel: coal	−0.631***	0.238	0.010
× Fuel: gas	−0.234*	0.118	0.051
× Fuel: oil	−0.481***	0.170	0.006
× Fuel: biomass	−0.122	0.111	0.275
× Capacity (std)	0.015	0.050	0.761
× Urban	0.071**	0.032	0.032

Note: Interaction model with allocation ratio interacted with fuel shares (continuous, summing to 1), standardized capacity, and urban indicator. Baseline treatment effect represents hypothetical facility with zero fuel shares, zero capacity, and rural location. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Fuel composition. The interaction model confirms that treatment effects are strongest for facilities using carbon-intensive fuels. The coal interaction (−0.631, $p = 0.010$) indicates that a facility with 100% coal share would experience a treatment effect of approximately $-0.046 - 0.631 = -0.68$, compared to $-0.046 - 0.234 = -0.28$ for a 100% gas facility. The biomass interaction is small and not statistically significant, consistent with biomass receiving preferential treatment under EU ETS carbon accounting [11].

Capacity. The capacity interaction is small and not significant (0.015, $p = 0.761$), indicating that treatment effects do not systematically vary with facility size after controlling for fuel composition.

Urbanization. The urban interaction is positive and significant (0.071, $p = 0.032$), indicating that urban facilities exhibit smaller emission reductions than rural facilities for a given change in allocation ratio. At mean urbanization (60% urban), this implies an attenuation of approximately 0.04 units relative to a fully rural facility.

For satellite NOx at the permissive detection limit (DL $\geq$ 0.03), the interaction model reveals significant effects for capacity and urbanization:

- Capacity interaction: -0.0014^{***} (SE = 0.0003, $p < 0.001$), indicating that larger facilities show stronger NOx reductions per unit change in allocation ratio
- Urban interaction: 0.0045^{***} (SE = 0.0011, $p < 0.001$), indicating that urban facilities show weaker (or reversed) NOx effects

These patterns are consistent across PCA and PLS specifications, providing evidence that while the main NOx effect is not significant at the permissive detection limit, there is meaningful heterogeneity in treatment effects by facility characteristics.

5.4 Robustness Summary

The results exhibit several patterns supporting robustness:

Cross-outcome consistency. Both ETS CO_2 and satellite NOx (at the conservative detection limit) show negative treatment effects, consistent with the hypothesis that reduced allocation ratios induce emission reductions. The sign consistency across independent measurement systems—administrative reporting versus satellite remote sensing—provides mutual validation.

Embedding method stability. NOx estimates at DL $\geq$ 0.11 are nearly identical across PCA (-0.00282) and PLS (-0.00301) specifications, differing by only 7%. This stability suggests that results are not sensitive to the specific dimensionality reduction approach and mitigates concerns about regularization bias from the supervised PLS method.

Detection limit sensitivity. The sharp contrast between null results at DL $\geq$ 0.03 and significant results at DL $\geq$ 0.11 is consistent with the measurement properties of the Beirle flux-divergence method: at permissive thresholds, noise dominates signal for most observations. The emergence of significant effects at the conservative threshold, where satellite estimates are more reliable, suggests that the null results at permissive thresholds reflect measurement error attenuation rather than absence of a true effect.

Measurement quality robustness. I examine heterogeneity along two satellite measurement quality dimensions:

- **Facility interference**: Facilities with another ETS installation within 20 km—where the satellite measurement may capture cluster-level rather than single-source emissions—show significant negative effects ($\beta = -0.0030$, $p = 0.013$). This addresses the concern that spatial interference could bias results: the treatment

effect persists even for interfered facilities, and if anything, cluster-level measurement should attenuate effects by averaging over sources with different treatment intensities.

- **Statistical integration error**: Split-sample analysis by statistical uncertainty (rel_err_stat $< 30\%$ vs. $\geq 30\%$) tests whether results are driven by high-uncertainty observations. The inverse-variance weighting in main specifications already accounts for heteroskedasticity; heterogeneity analysis confirms that treatment effects are stable across uncertainty levels.

Heterogeneity patterns. The stronger CO_2 effects for coal-dominant facilities ($\beta = -0.98$), electricity-sector facilities ($\beta = -0.21$), and rural facilities ($\beta = -0.24$) are theoretically plausible and consistent with the structure of EU ETS incentives. Some NOx heterogeneity patterns are anomalous (positive urban and coal coefficients), likely reflecting measurement challenges in those subgroups rather than genuine effect heterogeneity.

6 Discussion

This study develops and applies a methodological framework for evaluating climate policy impacts using dual outcomes: verified EU ETS CO_2 emissions and satellite-derived NOx emission proxies. The empirical results reveal a robust negative relationship between allocation ratios and verified CO_2 emissions, with consistent but noisier effects for satellite-derived NOx that emerge only at conservative detection thresholds. Several aspects of these findings merit discussion.

6.1 Summary of Empirical Findings

The main empirical finding is that EU ETS policy stringency—measured by the allocation ratio—has a robust negative effect on verified CO_2 emissions. The estimated coefficient of -0.186 ($p < 0.001$) implies that a 10% shortfall position (allocation ratio $= 0.9$) is associated with approximately 1.9% lower verified emissions compared to full free allocation. This effect is economically meaningful: at the sample mean of 580 $ktCO_2/yr$, it corresponds to 11 $ktCO_2$ per 0.1 unit decrease in allocation ratio.

The satellite-derived NOx proxy provides supporting evidence at the conservative detection limit (DL $\geq$ 0.11 kg/s), with estimated effects of -0.003 kg/s per unit allocation ratio ($p = 0.001$–0.017). This represents approximately 1.7% reduction in NOx emissions per 10% shortfall, broadly consistent with the CO_2 finding. However, at the permissive detection limit (DL $\geq$ 0.03 kg/s), the NOx effect is statistically indistinguishable from zero, highlighting the importance of detection limit thresholds in satellite-based emission studies.

Heterogeneity analysis reveals that treatment effects are strongest for coal-dominant facilities ($\beta = -0.98$ for CO_2), electricity-sector facilities ($\beta = -0.21$), and rural facilities ($\beta = -0.24$). These patterns are consistent with the structure of EU ETS incentives: coal faces the highest carbon intensity and hence the largest marginal abatement incentive, while electricity generators have largely lost free allocation under Phase III/IV [11].

Robustness checks on satellite measurement quality provide additional confidence. Treatment effects persist for facilities with spatial interference from nearby ETS installations ($\beta = -0.0030$, $p = 0.013$), indicating that cluster-level measurement does not drive spurious results. Similarly, effects are stable across levels of statistical integration uncertainty, consistent with the inverse-variance weighting approach accounting for observation-level precision.

6.2 Interpretation of Estimates

The allocation ratio treatment variable has a natural interpretation in terms of policy stringency. The coefficient of -0.186 on log-transformed verified CO_2 emissions implies that each 0.1 unit decrease in the allocation ratio is associated with approximately 1.9% lower emissions (since $\exp(-0.186 \times 0.1) - 1 \approx -0.018$). The effect is identified from within-facility variation in allocation ratios over time, after controlling for region-year fixed effects.

For verified CO_2: The estimate of -0.186 is precisely estimated (SE = 0.030) and highly significant ($p < 0.001$). The 95% confidence interval $[-0.246, -0.127]$ excludes zero and implies emission reductions of 1.3–2.5% per 0.1 unit shortfall. This estimate is conservative in the sense that any remaining endogeneity from dispatch variation would likely attenuate the coefficient toward zero.

For satellite-derived NOx: The estimates at the conservative detection limit (-0.00282 for PCA, -0.00301 for PLS) are significant at conventional levels ($p = 0.017$ and $p = 0.001$, respectively). The magnitude implies 0.3 g/s lower NOx per 0.1 unit decrease in allocation ratio. At the sample mean of 0.18 kg/s for above-threshold observations, this represents a 1.7% reduction—remarkably close to the 1.9% CO_2 reduction. However, the confidence intervals are wide due to small sample size (N = 140).

Cross-validation: The directional agreement between CO_2 and NOx estimates at the conservative detection limit supports the hypothesis that both outcomes are capturing genuine policy effects. The similar percentage magnitudes (1.9% vs. 1.7%) suggest that carbon pricing operates primarily through reduced combustion intensity rather than through emission factor changes (e.g., end-of-pipe controls). This pattern is consistent with the dominant mechanism being reduced output or fuel switching rather than pollution control investments.

The null results at the permissive detection limit illustrate the importance of signal-to-noise considerations in satellite-based emission studies. At DL $\geq$ 0.03 kg/s, the sample mean (0.065 kg/s) is below the standard detection threshold, meaning most observations are dominated by measurement noise. Classical measurement error attenuates treatment effects toward zero, explaining the null findings. This pattern—significant effects emerging only above detection thresholds—is precisely what physical measurement theory would predict.

6.3 Methodological Contributions: ML-Derived Features in Causal Inference

This study contributes to a growing literature on incorporating machine learning-derived features into causal inference frameworks [5–7]. Two aspects merit particular discussion.

Geospatial foundation model embeddings as controls. The use of AlphaEarth embeddings demonstrates that pre-trained geospatial representations can serve as effective high-dimensional controls in panel settings. The key assumption—that embeddings capture confounders affecting both policy exposure and air quality outcomes—is plausible given that they encode land use, infrastructure, and climate patterns that correlate with both industrial activity and pollution dispersion. Future work should investigate the conditions under which learned representations provide valid confounding adjustment.

Network-based clustering from external models. The use of PyPSA-Eur power system clusters for heterogeneity analysis represents a novel application of model-derived features. Due to sample attrition and the inclusion of non-electricity plants, these clusters are used only for heterogeneity analysis, not for identification or inference. The approach demonstrates that clusters derived from transmission network topology [9]—which group facilities facing correlated prices, dispatch patterns, and demand shocks—align with heterogeneous policy responses. This proof-of-concept could be extended to other networked industries where external models of network structure are available.

6.4 Implications for Policy Monitoring and Enforcement

The dual-outcome approach has practical implications for ETS monitoring and enforcement. The directional agreement between administrative CO_2 and satellite NOx provides a consistency check: if the two measures systematically diverge for a facility, this could flag potential data quality issues or unusual operational patterns. The satellite outcome is not accurate enough to serve as a standalone verification tool, but the NOx signal at high detection limits can corroborate administrative reporting.

More broadly, as satellite instruments improve in resolution and as methods like Beirle flux-divergence become refined (e.g., PAL product, stack-height winds), satellite-derived NOx could evolve into an independent verification channel. The framework developed here—linking administrative compliance data with satellite observations at the facility level—provides the methodological foundation for such applications.

6.5 Limitations and Future Work

6.5.1 Identification Concerns

Several potential threats to identification remain.

Operational confounding. Facilities may respond to high carbon prices by adjusting operations in ways not fully captured by the control variables, such as investing in pollutant abatement technologies. To the extent that these operational

responses are the mechanism through which carbon pricing affects air quality, this is not problematic—it is the causal effect of interest. However, if operational changes are driven by other factors correlated with allocation ratios (e.g., electricity prices), bias may result.

Spillovers. If carbon pricing induces substitution across facilities (e.g., shifting generation from high-cost to low-cost plants), the stable unit treatment value assumption (SUTVA) may be violated. The region×year fixed effects specification partially addresses this by absorbing regional substitution patterns. Future work can include power system observables from `PyPSA-EUR` to account for this.

6.5.2 Satellite NOx Proxy Limitations

The satellite-derived NOx estimates carry structural uncertainty from multiple sources. Total typical relative uncertainty is ∼35–45%, dominated by: (i) use of OFFL L3 instead of PAL NO_2 product (10–40% lower TVCDs); (ii) 50% uncertainty in lifetime correction; (iii) simplified NOx/NO_2 scaling (±7%); and (iv) unmodeled AMF and plume height corrections (±10% each).

Key design choices:

- **Detection limits**: The permissive threshold (0.03 kg/s) is methodologically informative but not reliable for causal claims; all substantive NOx conclusions use the conservative threshold (0.11 kg/s).
- **Spatial interference**: For facilities with another ETS facility within 20 km, the satellite outcome reflects cluster-level rather than single-facility emissions. Heterogeneity analysis confirms that treatment effects persist for interfered facilities ($\beta = -0.0030$, $p = 0.013$), suggesting that spatial interference does not drive spurious results.
- **Statistical integration error**: Observations with statistical integration error $\geq$ 30% are not excluded from main specifications; instead, inverse-variance weighting accounts for heteroskedasticity. Heterogeneity analysis by uncertainty level confirms treatment effect stability.
- **Skipping automatic identification**: I skip Beirle's point-source identification algorithm because I have known ETS/LCP coordinates. This is explicitly endorsed by the authors, but I apply simplified significance filters to compensate.

These choices primarily increase noise and attenuation bias, not spurious detection. The satellite outcome remains a *physically grounded but noisy proxy*; verified CO_2 remains the primary outcome for causal inference. The NOx results are best interpreted as corroborative evidence for the CO_2 findings and as a proof-of-concept for satellite-based monitoring.

6.5.3 Data Limitations

Sample attrition. The requirement for valid linkage across three independent data sources (LCP registry, EU Registry crosswalk, EUTL compliance data) reduces the initial universe of 3,405 LCP plants to 521 facilities (15.3% retention). The satellite NOx outcome has additional attrition from detection limits and observation coverage

requirements. This attrition reduces statistical power and may introduce selection bias if the matched sample differs systematically from the broader LCP population.

Generalizability. The sample is restricted to large facilities (LCPs), with sufficient emissions for satellite detection (0.11 kg/s NOx conservative threshold) in the case of the NOx outcome, limiting generalizability to smaller sources. The annual temporal resolution may miss short-run dynamics such as seasonal fuel switching or within-year operational adjustments.

Heterogeneous satellite observation coverage. Different facilities have different numbers of valid observation days per year (typically 60–80 out of ∼180 days with TROPOMI coverage) due to cloud cover, wind speed filtering (≥2 m/s requirement), and satellite orbit patterns. For panel regressions, this concern is mitigated if: (i) observation selection is driven by weather, which is exogenous to treatment; (ii) the selection mechanism is stable within-facility over time; and (iii) year fixed effects absorb common temporal patterns.

UK exclusion. The EU ETS registry data does not include UK installations following Brexit. UK large combustion plants—which represented a significant share of EU ETS-regulated capacity prior to 2021—are excluded. Future work could extend this framework to include UK facilities by obtaining compliance data from the UK ETS registry [28].

6.5.4 Future Work

Event-study analysis. The Callaway-Sant'Anna [10] estimator would provide a valuable robustness check through event-study plots and formal pre-trend tests. However, as discussed in Section 4.4, this approach is infeasible with the current panel (2018–2023) because 84.5% of ever-treated facilities were already treated in the first panel year. Extending the panel backward to include EU ETS Phase 3 (2013–2017) would enable event-study analysis with proper pre-treatment observations.

Power system integration. Incorporating dispatch and generation data from `PyPSA-EUR` would enable direct control for facility-level utilization, addressing the bad-controls trade-off discussed in Section 4.2.

Extension to other pollutants. The dual-outcome framework could be extended to methane point sources using TROPOMI CH_4, enabling comprehensive evaluation of climate and air quality co-benefits.

7 Conclusion

This study develops and demonstrates a novel framework for comprehensively evaluating climate policy impacts using dual emission outcomes. By linking administrative EU ETS compliance data with satellite-derived NOx emission proxies constructed via the Beirle-style flux-divergence method, I construct a facility-level panel that enables causal inference on how carbon market stringency affects both verified CO_2 emissions and satellite-observable combustion co-pollutants.

The study makes three methodological contributions. The first two follow a recent trend in causal inference toward incorporating machine learning-derived features to

address high-dimensional confounding [5–7]; the third adapts an atmospheric physics method for panel econometrics.

First, I demonstrate the integration of geospatial foundation model embeddings as high-dimensional controls in panel-based climate monitoring studies. Using Google AlphaEarth [8]—64-dimensional representations learned from multi-source satellite imagery, climate reanalysis, and geotagged text—I capture between-unit heterogeneity arising from local geographic, infrastructural, and climate context in a data-efficient manner. This extends prior work on learned embeddings for causal inference [6] from text to the geospatial domain, and is particularly relevant for difference-in-differences settings where high-dimensional spatial confounders may violate parallel trends if left uncontrolled [7].

Second, I use Eurostat NUTS2 regions for spatial clustering in both fixed effects structure and inference, with PyPSA-Eur power system clusters [9] used for electricity-sector heterogeneity analysis. Due to sample size limitations and the inclusion of non-electricity facilities, PyPSA-Eur clusters are *not* used for identification or inference—only for exploring treatment effect heterogeneity across network-defined regions. This represents a proof-of-concept that network-derived clusters (k-means on transmission network topology) align with heterogeneous policy responses in electricity systems. The approach is grounded in recent theoretical foundations showing that valid cluster-robust inference requires low-conductance clusters [18], which network-derived clusters satisfy by construction.

Third, for the satellite outcome, I implement a simplified Beirle-style flux-divergence method that provides physically grounded NOx emission estimates at the facility level. By computing the advection—the wind-aligned spatial derivative of NO_2 column density—and integrating over a 15 km disc with lifetime correction, this approach provides facility-specific NOx estimates suitable for panel econometric analysis. This methodology is grounded in the continuity equation and follows the approach of Beirle et al. (2019, 2021, 2023) [1–3], which has been validated against reported emissions from regulatory agencies.

Empirical findings. Applying this framework to 521 EU ETS-regulated large combustion plants across 82 NUTS2 regions (2018–2023), I find that carbon market stringency has a robust negative effect on verified CO_2 emissions. A 10% allocation shortfall (allocation ratio = 0.9 vs. 1.0) is associated with approximately 1.9% lower verified emissions ($\beta = -0.186$, $p < 0.001$). This effect is concentrated among coal-dominant facilities ($\beta = -0.98$) and electricity generators ($\beta = -0.21$), consistent with the structure of EU ETS incentives.

The satellite-derived NOx proxy provides supporting evidence at the conservative detection limit: facilities above 0.11 kg/s NOx show significant negative effects of similar magnitude ($\beta \approx -0.003$, 1.7% per 10% shortfall, $p < 0.01$). The directional agreement between administrative CO_2 and satellite NOx—outcomes measured through entirely independent systems—strengthens confidence that both are capturing genuine policy effects. Robustness checks confirm that effects persist for facilities with spatial interference from nearby ETS installations, and are stable across levels of statistical measurement uncertainty. However, the null results at permissive

detection limits highlight the continued importance of signal-to-noise considerations in satellite-based emission monitoring.

The broader contribution of this work is demonstrating that combining administrative emissions records with satellite-derived proxies, along with ML-derived controls and network-informed inference, can provide comprehensive evaluation of climate policy impacts at the individual emitter level. The dual-outcome approach offers several advantages: (i) verified CO_2 provides the gold standard for measuring policy effects on greenhouse gas output; (ii) satellite-derived NOx provides an independent check on combustion activity and enables testing co-benefit hypotheses; and (iii) agreement between outcomes provides cross-validation that both measures are capturing genuine policy effects.

As satellite instruments improve in resolution and retrieval accuracy, and as methods like the Beirle flux-divergence approach become more refined, this framework could enable near-real-time monitoring of both carbon and co-pollutant emissions from regulated facilities. Future work could extend this framework to methane point sources (using TROPOMI CH_4), investigate heterogeneity across plant types and regulatory contexts, and further develop the theoretical foundations for using learned representations and model-derived clusters in causal inference.

Acknowledgements.

Declarations

- **Data availability**: EEA Large Combustion Plant data available from the European Environment Agency Industrial Emissions Portal. EU ETS data available from the European Union Transaction Log. TROPOMI data available via Google Earth Engine. ERA5-Land data available from the Copernicus Climate Data Store.
- **Code availability**: Full data processing and analysis code is available at https://github.com/arnava13/Masters-Thesis
- **Use of Generative AI**: Generative AI was used for programming, researching and writing this paper. Project direction was driven by me, and I manually verified and refactored all code, sources, citations, equations and theoretical assertions.

Appendix A Data Pipeline Details

A.1 ID Normalization for ETS Linking

Linking LCP plants to ETS installations requires normalizing identifiers from different sources. EU Registry identifiers follow patterns such as `FR000000000210535` (padded numeric) or `FR-new-07101261` (new format). Pyeutl installation IDs follow the format `AT_200165` (country code underscore numeric).

The normalization procedure:

1. Extract country code (first 2 characters)
2. Extract all numeric substrings
3. Select longest numeric substring, strip leading zeros
4. Combine as `CC_NNN` format

This procedure successfully matches 799 of 932 facilities (85.7%) to ETS installations.

A.2 Electricity Sector Classification

Electricity-sector facilities are identified using EU ETS activity codes from the EUTL database, as defined in Directive 2003/87/EC Annex I [29, 30]. Activity codes changed between EU ETS phases:

- **Phases 1–2 (2005–2012)**: Activity Code 1 = "Combustion installations with a rated thermal input exceeding 20 MW"
- **Phase 3+ (2013–present)**: Activity Code 20 = "Combustion of fuels"

Strictly speaking, these activity codes identify *combustion installations* broadly—including power plants, combined heat and power (CHP), industrial boilers, and district heating—rather than electricity generators specifically. However, for the Large Combustion Plant (LCP) registry used in this study, the sample predominantly comprises electricity-generating facilities. A facility is classified as electricity-sector if it has *any* installation linked to activity codes 1 or 20. This classification is used for electricity-sector heterogeneity analysis employing PyPSA-Eur power system clusters.

A.3 Fuel Type Classification

Raw LCP fuel types are mapped to standardized categories:

- **Gas**: NaturalGas, NG, Gas
- **Coal**: Coal, Lignite, PC, BIT, SUB, ANT
- **Oil**: LiquidFuels, DFO, RFO, KER
- **Biomass**: Biomass, WDL, WDS, AB
- **Other Gas**: OtherGases, OBG

Fuel types used by fewer than 10% of facility-years (Other Solid, Peat) are dropped, shares renormalized, and facilities with no remaining fuel coverage are excluded from the sample (although there were no such facilities).

Appendix B Sample Attrition

Table B1 summarizes the sample attrition through each processing step. The most significant losses occur at the ETS linkage stage (44% of plants lack matched ETS identifiers in the EU Registry crosswalk) and the requirement for matched ETS compliance data with valid allocation ratios.

References

[1] Beirle, S. *et al.* Pinpointing nitrogen oxide emissions from space. *Science Advances* **5**, eaax9800 (2019).

Table B1 Sample Attrition Through Data Processing Pipeline

Processing Step	Plants/Facilities	Lost	Retained %
Plant-Level Processing			
Initial LCP registry (≥50 MW thermal)	3,405 plants	—	100%
With complete capacity + fuel data (2018–2023)	2,821	584	82.8%
With ETS linkage (via EU Registry crosswalk)	1,580	1,241	46.4%
Facility-Level Processing (after 500m spatial clustering)			
After spatial clustering	932 facilities	—	—
With matched ETS compliance data	608	324	65.2%
With ≥3 years complete data	521	87	55.9%
Base analysis panel (ETS CO_2 outcome)	**521 facilities** **2,819 fac-years**		**15.3%** **of initial plants**
Additional NOx Outcome Filters			
With satellite data (≥20 valid days/year)	291	230	55.9%
With ≤50% total uncertainty	291	0	55.9%
Satellite NOx panel	**291 facilities** **1,213 fac-years**		**8.5%** **of initial plants**

Note: LCP registry includes only plants with rated thermal input ≥50 MW. ETS linkage uses normalized identifier matching between EU Registry and EUTL compliance data. The satellite NOx panel filters on observation coverage (≥20 valid days/year) and total uncertainty (≤50%). Detection limit (0.11 kg/s) is not used as a sample filter but is reported as a quality indicator.

[2] Beirle, S. *et al.* Catalog of NOx emissions from point sources as derived from the divergence of the NO_2 flux for TROPOMI. *Earth System Science Data* **13**, 2995–3012 (2021).

[3] Beirle, S., Borger, C., Jost, A. & Wagner, T. Improved catalog of NOx point source emissions (version 2). *Earth System Science Data* **15**, 3051–3073 (2023). Key methodological reference for flux-divergence approach, lifetime correction, and NOx/NO_2 scaling.

[4] Vandyck, T. *et al.* Air quality co-benefits for human health and agriculture counterbalance costs to meet Paris Agreement pledges. *Nature Communications* **9**, 4939 (2018).

[5] Chernozhukov, V. *et al.* Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* **21**, C1–C68 (2018). Key reference for regularization bias in ML-based causal inference and sample-splitting strategies.

[6] Veitch, V., Sridhar, D. & Blei, D. M. Adapting text embeddings for causal inference (2019). ArXiv:1905.12741, arXiv:1905.12741.

[7] Zimmert, M. Efficient difference-in-differences estimation with high-dimensional common trend confounding (2018). ArXiv:1809.01643, arXiv:1809.01643.

[8] Rolf, E. *et al.* AlphaEarth foundations: An embedding field model for accurate and efficient global mapping from sparse label data (2025). ArXiv:2507.22291, arXiv:2507.22291.

[9] Hörsch, J., Hofmann, F., Schlachtberger, D. & Brown, T. PyPSA-Eur: An open optimisation model of the European transmission system. *Energy Strategy Reviews* **22**, 207–215 (2018). Network topology from ENTSO-E; clustering reduces computational complexity while preserving electrical characteristics.

[10] Callaway, B. & Sant'Anna, P. H. C. Difference-in-differences with multiple time periods. *Journal of Econometrics* **225**, 200–230 (2021).

[11] Ellerman, A. D., Marcantonini, C. & Zaklan, A. *The European Union Emissions Trading System: Ten Years and Counting* Vol. 10 (Review of Environmental Economics and Policy, 2016).

[12] Beirle, S. & Wagner, T. A new method for estimating megacity NOx emissions and lifetimes from satellite observations. *Atmospheric Measurement Techniques* **17**, 3439–3453 (2024).

[13] Jiao, L., Liu, Y. & Zou, B. Satellite verification of ultra-low emission reduction effect of coal-fired power plants. *Atmospheric Pollution Research* **11**, 1839–1847 (2020).

[14] Castellanos, P. & Boersma, K. F. Reductions in nitrogen oxides over Europe driven by environmental policy and economic recession. *Scientific Reports* **2**, 265 (2012).

[15] Fioletov, V. *et al.* Quantifying urban, industrial, and background changes in NO_2 during the COVID-19 lockdown period based on TROPOMI satellite observations. *Atmospheric Chemistry and Physics* **22**, 4201–4236 (2022).

[16] Goodman-Bacon, A. Difference-in-differences with variation in treatment timing. *Journal of Econometrics* **225**, 254–277 (2021).

[17] Sun, L. & Abraham, S. Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics* **225**, 175–199 (2021).

[18] Kojevnikov, D., Marmer, V. & Song, K. Network cluster-robust inference. *Econometrica* **91**, 641–667 (2023).

[19] NASA Jet Propulsion Laboratory. NASA SRTM Digital Elevation 30 m (SRTMGL1 v003). NASA LP DAAC, USGS/Earth Resources Observation and Science (EROS) Center (2013). URL https://developers.google.com/earth-engine/datasets/catalog/USGS_SRTMGL1_003. Accessed via Google Earth Engine (dataset ID USGS/SRTMGL1_003).

[20] Farr, T. G. *et al.* The shuttle radar topography mission. *Reviews of Geophysics* **45**, RG2004 (2007).

[21] Schiavina, M., Melchiorri, M. & Pesaresi, M. GHS-SMOD R2023A — GHS settlement layers, application of the degree of urbanisation methodology (stage I) to GHS-POP R2023A and GHS-BUILT-S R2023A, multitemporal (1975–2030). European Commission, Joint Research Centre (JRC) (2023). URL https://developers.google.com/earth-engine/datasets/catalog/JRC_GHSL_P2023A_GHS_SMOD_V2-0. Global Human Settlement Layer Degree of Urbanisation. SMOD classes: 10=Water, 11-13=Rural, 21=Suburban, 22-23=Urban cluster, 30=Urban centre. Accessed via Google Earth Engine.

[22] Wikipedia contributors. Latitude — Wikipedia, the free encyclopedia (2024). URL https://en.wikipedia.org/wiki/Latitude#Meridian_distance_on_the_ellipsoid. Section: Meridian distance on the ellipsoid. WGS84 series expansion accurate to 0.01 m/degree.

[23] Lange, K., Richter, A. & Burrows, J. P. Variability of nitrogen oxide emission fluxes and lifetimes estimated from Sentinel-5P TROPOMI observations. *Atmospheric Chemistry and Physics* **22**, 2745–2767 (2022). Latitude-dependent NOx lifetime parameterization used in Beirle v2.

[24] NIST. Nitrogen dioxide (NO_2). NIST Chemistry WebBook, SRD 69 (2023). URL https://webbook.nist.gov/cgi/cbook.cgi?ID=10102-44-0. CAS 10102-44-0, Molar mass 46.0055 g/mol.

[25] European Commission. ETS revision: No change to deadline to surrender allowances in 2023. Directorate-General for Climate Action (2023). URL https://climate.ec.europa.eu/news-your-voice/news/ets-revision-no-change-deadline-surrender-allowances-2023-2023-01-30_en. Accessed December 2024. Confirms compliance calendar: free allocation by 28 February, surrender by 30 April.

[26] Cinelli, C., Forney, A. & Pearl, J. A crash course in good and bad controls. *Sociological Methods & Research* (2022). Defines bad controls as variables that, when conditioned on, introduce bias. Relevant to supervised dimensionality reduction where outcome information leaks into covariates.

[27] Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual (2024). URL https://www.gurobi.com.

[28] UK Government. UK Emissions Trading Scheme (UK ETS): A policy overview. UK Government Policy Paper (2024). URL https://www.gov.uk/government/publications/uk-emissions-trading-scheme-uk-ets-policy-overview/uk-emissions-trading-scheme-uk-ets-a-policy-overview. Accessed December 2024.

[29] European Parliament and Council. Directive 2003/87/EC establishing a scheme for greenhouse gas emission allowance trading within the Community (2003). URL https://eur-lex.europa.eu/eli/dir/2003/87/oj. Annex I defines activity codes: Code 1 (Phases 1–2) = 'Combustion installations with rated thermal input exceeding 20 MW'; Code 20 (Phase 3+) = 'Combustion of fuels'. Consolidated version at https://eur-lex.europa.eu/eli/dir/2003/87/2024-03-01.

[30] European Environment Agency. EU ETS data viewer: User manual and background note. Technical Document, European Environment Agency (2021). URL https://www.eea.europa.eu/data-and-maps/data/european-union-emissions-trading-scheme-12/eu-ets-background-note. Table 6-1 provides activity codes used in EUTL database; codes 1, 20 identify combustion/-electricity sector installations.