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

This study develops a novel framework for evaluating climate policy impacts using two complementary emission outcomes. I investigate how European Union Emissions Trading System (EU ETS) policy stringency affects both installation-level verified CO_2 emissions and satellite-derived NOx emission proxies around major industrial emitters. The dual-outcome approach enables cross-validation: EU ETS verified emissions provide high-quality, installation-level measures of greenhouse gas output, while satellite-derived NOx estimates offer physically grounded proxies for combustion co-pollutants that can reveal co-benefits and potential under-reporting. The methodological framework makes three primary contributions. First, I demonstrate the integration of geospatial foundation model embeddings (Google AlphaEarth, 64 dimensions) as controls in panel-based climate monitoring studies, capturing between-unit heterogeneity arising from local geographic and climate context in a data-efficient manner that would be impractical to specify manually. Second, I introduce network-based clustering for inference: standard errors and region-by-time fixed effects use PyPSA-Eur power system clusters—k-means clusters computed on transmission network topology features from an external power system model—that group facilities facing correlated wholesale prices, dispatch patterns, and grid constraints. Third, for the satellite outcome, I implement a simplified Beirle-style flux-divergence method: using TROPOMI NO_2 tropospheric column densities and ERA5 winds, I compute the advection (wind-aligned spatial derivative) of NO_2 fields around each facility, integrate within a 15 km radius, and apply lifetime and NO_2-to-NOx corrections following Beirle et al. (2023). This approach is physically grounded in the continuity equation and specifically designed for power-plant-scale NOx

plumes. The analysis panel links 521 EU ETS-regulated large combustion plants across Europe (2018–2023) with regulatory emissions data, employing two-way fixed effects and Callaway-Sant'Anna difference-in-differences estimators. This work demonstrates a novel application of recent advancements in both machine-learning aided causal inference as well as remote-sensing derived pollutant outcomes into econometric panel analysis.

Keywords: Climate Policy, 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 two primary methodological contributions, both following a recent trend in causal inference toward incorporating machine learning-derived features to address high-dimensional confounding [5–7].

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 time fixed effects, as well as the Callaway and Sant'Anna [10] difference-in-differences estimator for staggered treatment timing.

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 four 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.

This study implements both traditional TWFE specifications (which remain valid under homogeneous treatment effects) and the Callaway-Sant'Anna estimator (which is robust to heterogeneity), allowing comparison of results under different identifying assumptions.

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 representations, learned for prediction tasks on large corpora, may 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 both fixed effects structure and clustered inference. 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 inference in policy evaluation.

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 in interpretation.

3.1.4 ERA5-Land Reanalysis

Hourly 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 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) [19]. 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)

Why Urbanization is Not a Regression Control.

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 [20]:

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

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 (peat, other solid fuels) are dropped, remaining shares are renormalized to sum to unity, and facilities with no remaining fuel coverage are excluded.

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.

Emissions Filter (ETS CO_2 Only).

For the ETS CO_2 outcome analysis, facilities are required to have at least one year with verified emissions $\geq$100 ktCO_2/yr to ensure sufficient policy exposure magnitude. This filter is applied *only* to the ETS CO_2 analysis panel, *not* to the satellite NOx panel. The satellite outcome has its own detection limits (0.03–0.11 kg/s NOx) based on the Beirle methodology, which are independent of CO_2 emissions. Applying a CO_2-based filter to the satellite panel would be methodologically incorrect, as facilities with low CO_2 emissions may still have detectable NOx signals (e.g., due to high NOx/CO_2 ratios from certain fuel types or combustion technologies).

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. This design choice is explicitly endorsed by the authors' statement and is conceptually appropriate: the identification algorithm is needed only when constructing a global catalog without prior knowledge of emission sources, not when applying the physically grounded quantification to facilities whose locations are already known.

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 (molecules/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 × 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 × 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 $0.0167 \times \exp(-0.575/\cos(\mathrm{SZA}))$ s^{-1}), 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 SRTM DEM), 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):

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

where A^* has units mol/(m$^2\cdot$ 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. [21] via Beirle et al. Eq. (10):

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

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) [22]. 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:

- **Statistical error** (Sect. 3.12.2): I approximate using the standard error of the temporal mean of daily integrated emissions, rather than Beirle's per-pixel SE propagation. This is more conservative as it captures meteorological variability in addition to sampling noise, typically <10%. Facilities with statistical relative error ≥30% are flagged via `rel_err_stat_lt_0_3`.
- **Lifetime correction** (Sect. 3.12.1): 50% relative uncertainty in τ, propagated through $c_\tau = \exp(t_r/\tau)$ yielding $\sigma_{c_\tau}/c_\tau = \ln(c_\tau) \times 0.50$, typically 10–20% for $c_\tau \approx 1.4$.
- **NOx/NO_2 scaling** (Sect. 3.12.1): ±0.10 on 1.38 ratio, ∼7%.
- **AMF correction** (Sect. 3.12.1): *Unmodeled structural uncertainty*—I do not implement an explicit AMF correction. A 10% term is carried as a generic structural uncertainty following Beirle's error budget, representing potential bias rather than fitted variance.
- **Plume height** (Sect. 3.12.3): *Unmodeled structural uncertainty*—I do not implement plume-height-dependent wind interpolation. A 10% term represents the sensitivity to assumed height (500m vs 300m) as reported by Beirle.
- **Topographic correction** (Sect. 3.12.4): 33% uncertainty on $f = 1.5$, typically <2.5% for flat European terrain.
- **OFFL vs PAL product** (our addition): OFFL provides 10–40% lower TVCDs than PAL; a 25% structural uncertainty term is added to account for this systematic difference.

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

Uncertainty-based sample restriction. 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.

Detection limits and significance flags. 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).
- `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.

Main satellite regressions restrict to "significant" observations satisfying `above_dl_0_11` ∧ `rel_err_stat_lt_0_3`. Sensitivity analyses additionally exclude interfered facilities or relax to the permissive detection limit.

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 1 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 that encourage embeddings 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.

Since the embeddings are derived from satellite imagery aggregated over time, they are treated as static facility-level controls. The 64 dimensions are included directly without dimensionality reduction, as the panel fixed effects structure provides regularization against overfitting. This represents 64 bytes per location—a highly compressed representation of the local geographic and climate context that would require hundreds of manually-specified variables to approximate. As discussed in Section 4.2.6, these 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.

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.

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.

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.

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

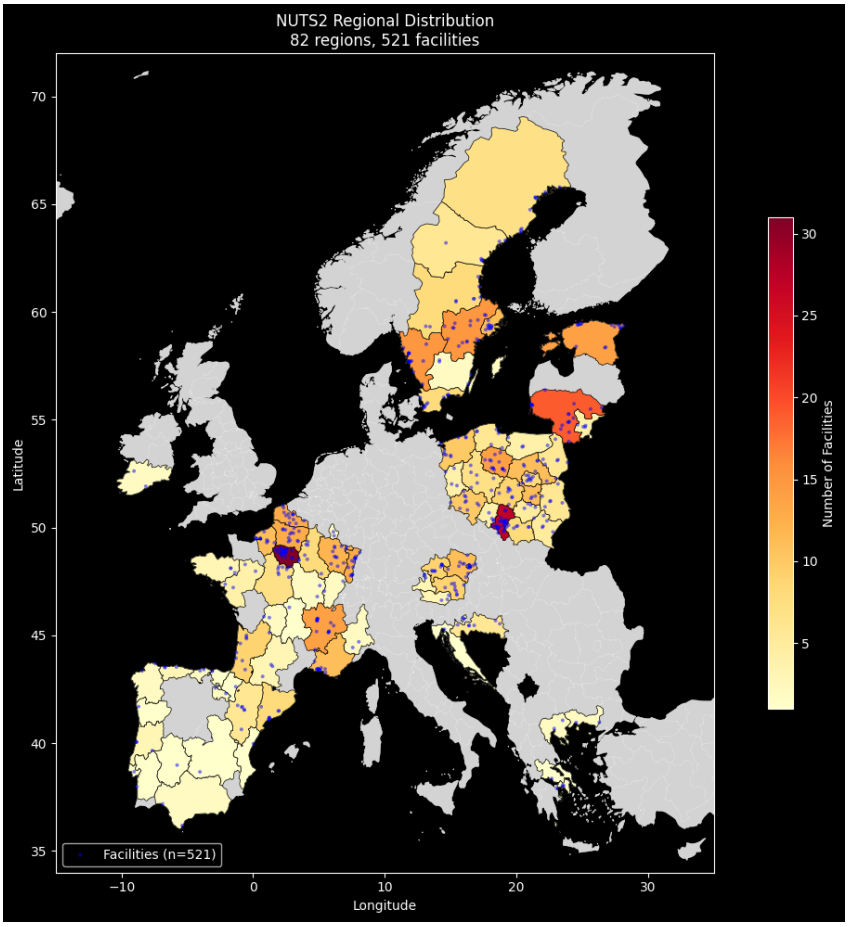

Fig. 1 Geographic distribution of 521 facilities across 82 NUTS2 regions. Color intensity indicates facility count per region. Blue points mark individual facility locations.

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.

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.

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

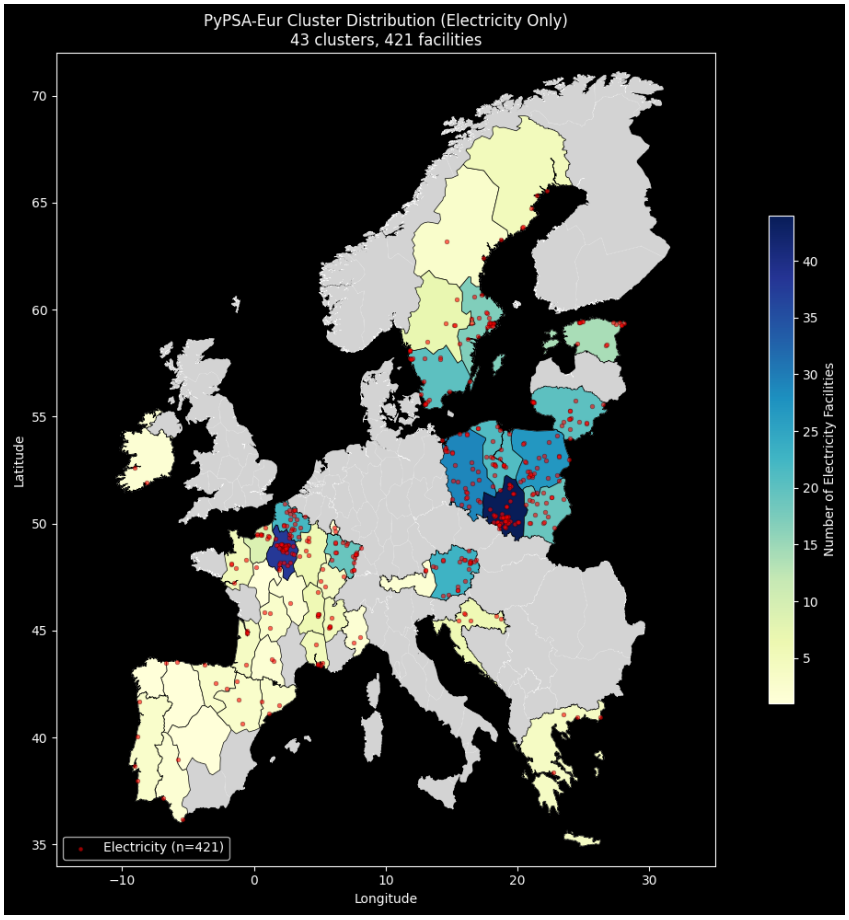

Fig. 2 Distribution of 421 electricity-generating facilities across 43 PyPSA-Eur power system clusters. Clusters are derived from k-means clustering on transmission network topology, grouping facilities facing correlated wholesale prices and grid constraints.

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.

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.

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

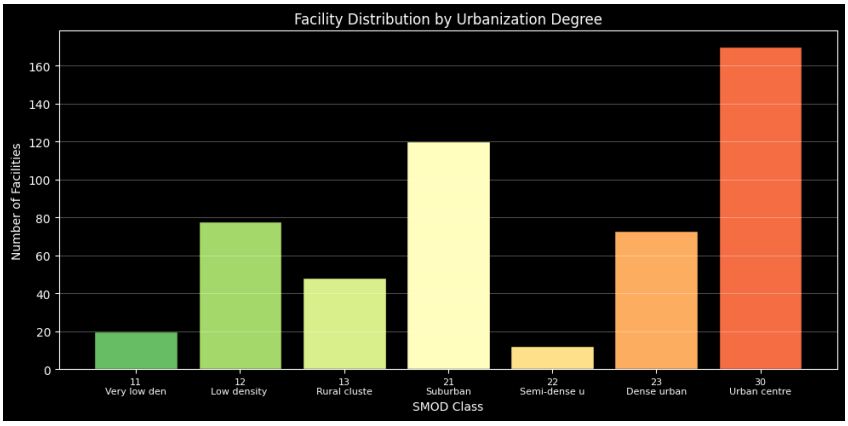

Fig. 3 Distribution of facilities by urbanization degree (GHSL-SMOD classification). Categories range from very low density rural (11) to urban centers (30). The majority of facilities are in suburban (21) and urban center (30) locations.

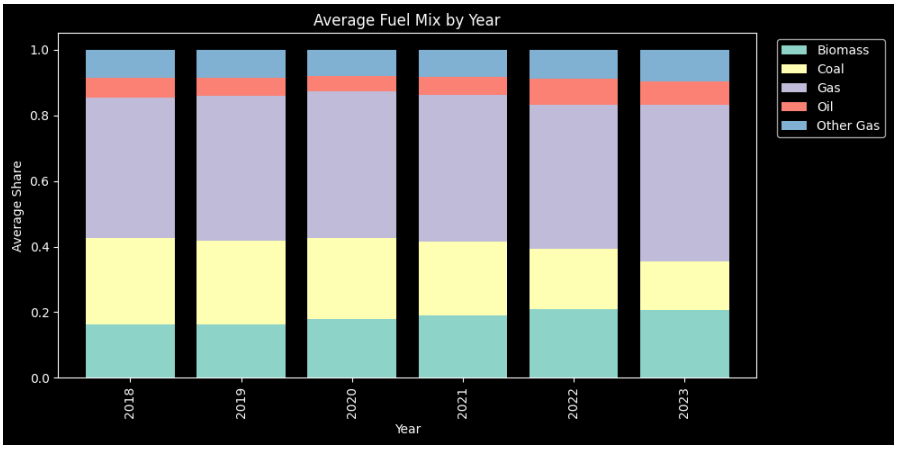

Fig. 4 Average fuel mix by year across all facilities. Gas (purple) dominates, with coal (tan) showing a modest decline over the sample period. Biomass (teal) and other gas (blue) shares increase slightly.

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 will necessitate 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

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).

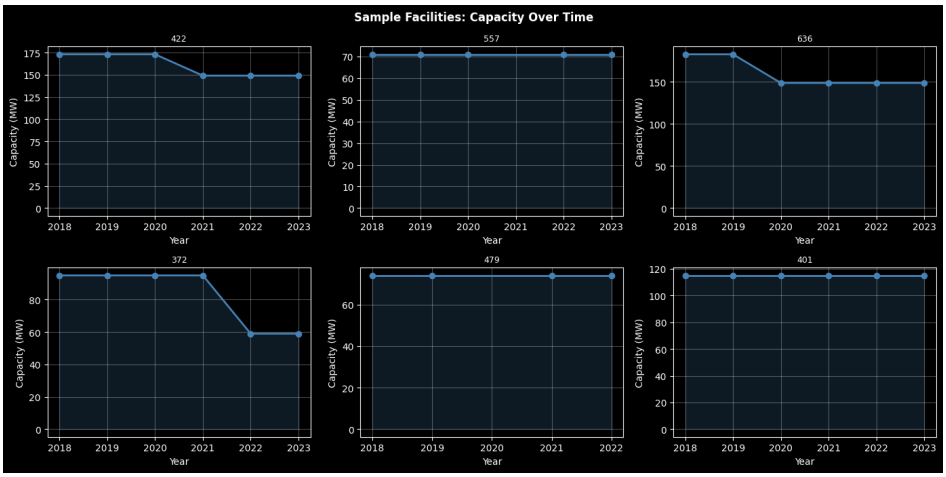

Fig. 6 Rated thermal capacity (MW) over time for six randomly sampled facilities. Most facilities exhibit stable capacity with occasional step changes.

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 (±50%), NO_2/NOx ratio uncertainty (±10%), and satellite product-related errors. The `nox_weight` variable ($= 1/\sigma^2_{\text{rel}}$) provides inverse-variance weights for regression analysis.

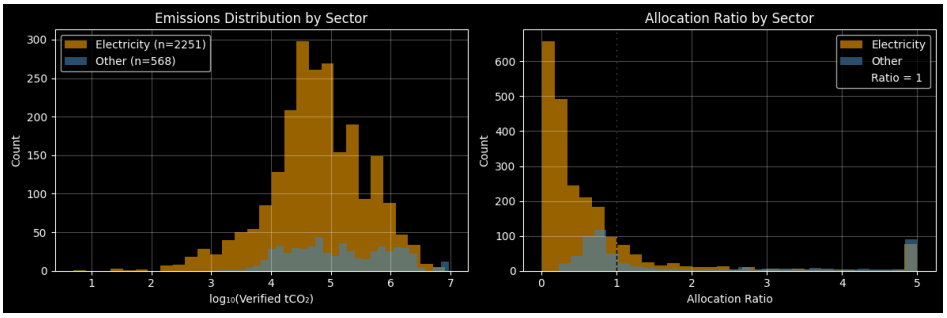

Fig. 7 Distribution of verified emissions (left) and allocation ratios (right) by sector. Electricity facilities (orange) have higher emissions but lower allocation ratios due to reduced free allocation under EU ETS Phase III/IV.

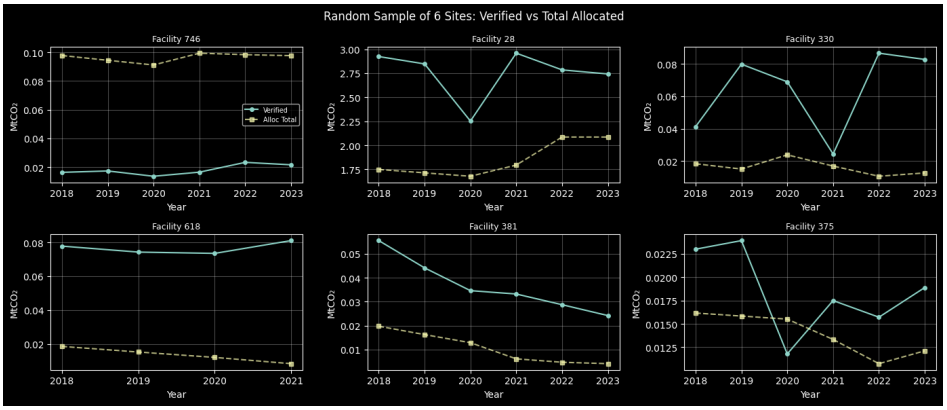

Fig. 8 Verified emissions (teal solid) versus free allocations (yellow dashed) for six randomly sampled facilities. The persistent gap above the allocation line indicates shortfall positions requiring market purchases.

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.

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

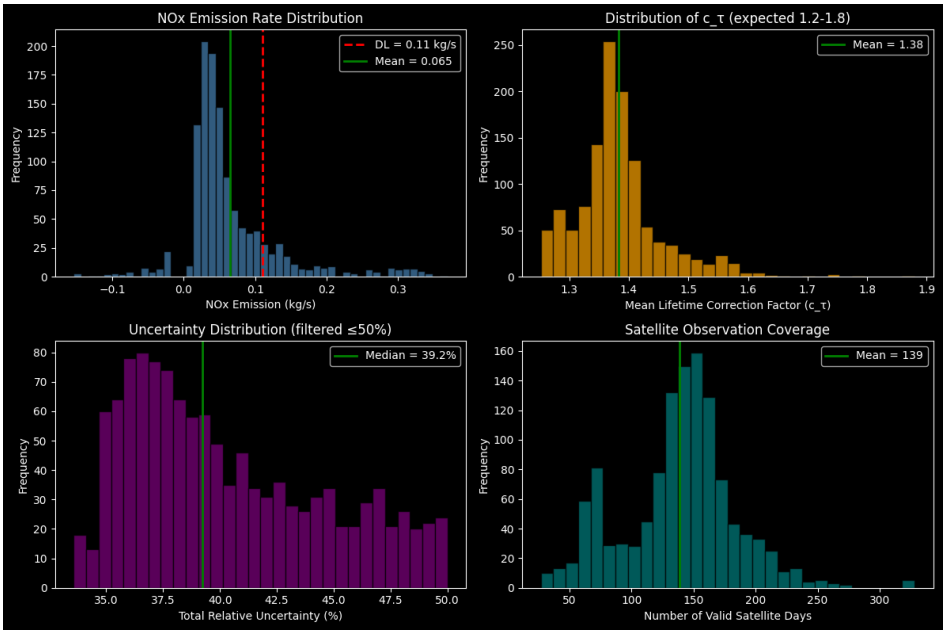

Fig. 9 Satellite-derived NOx emission characteristics for 291 facilities (1,213 facility-years). Top-left: NOx emission rate distribution with detection limit (0.11 kg/s) and mean (0.065 kg/s). Top-right: Lifetime correction factor distribution (mean 1.38). Bottom-left: Total relative uncertainty (median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

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 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.

Table 1 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.

- **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.

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.

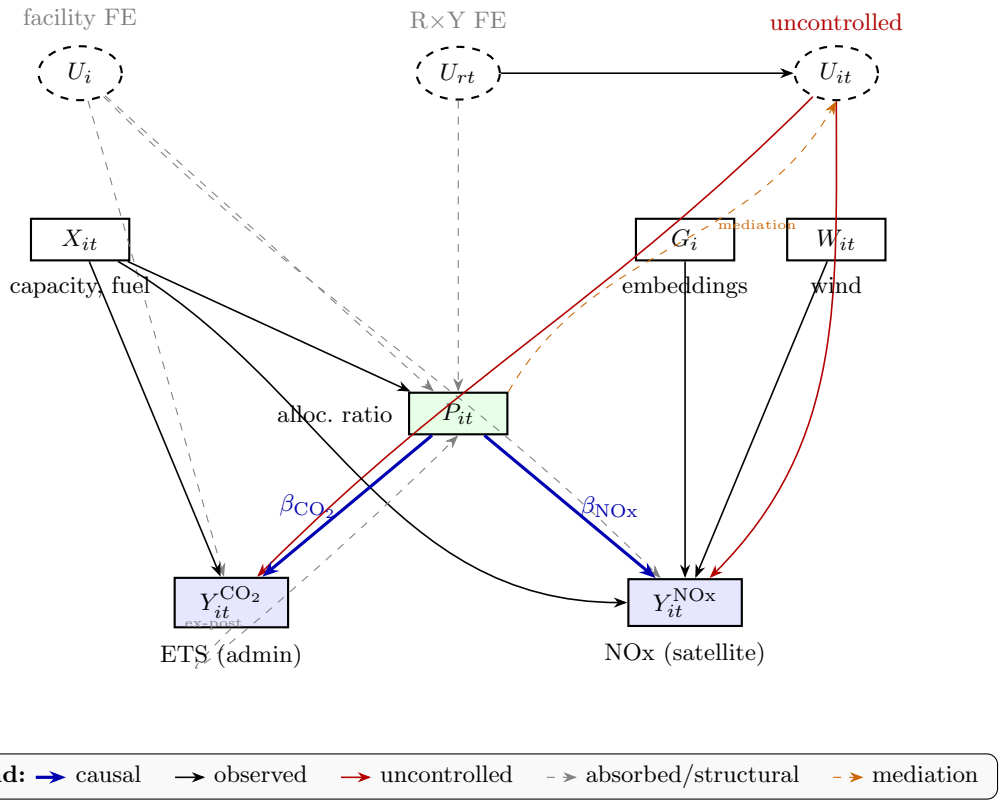

Fig. 10 Directed acyclic graph for dual-outcome causal inference. Treatment P_{it} (allocation ratio) affects both verified CO_2 and satellite NOx outcomes. Gray dashed arrows from U_i and U_{rt} are absorbed by facility and region×year fixed effects, respectively. Red arrows from U_{it} indicate residual confounding from time-varying unobservables (dispatch, maintenance) that we 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.

4.2.1 Why We Control for 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 Why We Do Not Control for U_{it} (Plant-Level Time-Varying Unobservables)

Plant-level time-varying unobservables (U_{it}) include dispatch/utilization, maintenance status, operational efficiency changes, and idiosyncratic output demand. These variables—particularly dispatch—present a "bad control" problem because dispatch is simultaneously:

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

Controlling for U_{it} (or proxies such as generation data) would absorb both effects. The confounding component should be removed, but blocking the mediation pathway would attenuate the true policy effect toward zero. Since we cannot empirically separate these components without strong assumptions about the dispatch mechanism, we do not control for U_{it} directly. Instead, Region×Year FE absorbs the common (regional) component of U_{it}—since dispatch responds primarily to regional demand and fuel prices—leaving only facility-specific deviations as residual confounding. These facility-specific deviations are plausibly second-order and orthogonal to the allocation ratio conditional on capacity and fuel mix controls.

4.2.3 Why Not Facility×Year Fixed Effects?

Facility×year fixed effects (α_{it}) would absorb *all* within-facility-year variation. Since treatment varies at the facility-year level, this leaves no variation to identify β. More subtly, interactive fixed effects models (which estimate facility-specific responses to common time factors) risk similar problems: if the estimated factor loadings correlate with how allocations were assigned, conditioning on them may open backdoor paths through the allocation mechanism or block mediation pathways.

4.2.4 Addressing Simultaneity in the Allocation Ratio

The allocation ratio $R_{it} = A_{it}/V_{it}$ involves current-year verified emissions V_{it} in the denominator, creating apparent simultaneity. However, the **temporal structure of EU ETS compliance** eliminates this concern. The annual compliance cycle proceeds as follows:

1. **Free allocation granted** (28 February of year t): Competent authorities issue free allowances A_{it} to each installation based on predetermined benchmarks [23].

2. **Emissions occur** (throughout year t): Facilities make operational decisions—dispatch, fuel choice, maintenance—knowing their allocation A_{it}.
3. **Emissions verified and reported** (31 March of year $t+1$): Accredited verifiers audit emissions V_{it}.
4. **Allowances surrendered** (30 April of year $t+1$): Facilities surrender allowances equal to verified emissions.

The key insight is that A_{it} is **known before** emissions decisions are made. Facilities observe their allocation at the start of the year and adjust operations accordingly. The causal direction is unambiguous: $A_{it} \rightarrow$ decisions $\rightarrow V_{it}$. The dashed arrow from $Y_{it}^{CO_2}$ to P_{it} in Figure 10 represents the ex-post calculation of the ratio, not reverse causation within the compliance period.

Within the econometric design, Region×Year FE absorbs common drivers of dispatch variation (carbon prices, regional demand, fuel prices). The remaining cross-facility variation in R_{it} within a region-year reflects primarily differences in predetermined allocations A_{it} rather than differences in current dispatch.

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.

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 idiosyncratic demand for a specific plant's output. These are plausibly second-order and unlikely to systematically correlate with allocation ratios conditional on our controls. Future work incorporating plant-level generation data could address this directly; a discussion of economic dispatch and power system optimization 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 [24]: 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. 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 Treatment Definitions

4.3.1 Continuous Treatment

The primary treatment variable is the allocation ratio R_{it} (Equation 9). Lower values indicate greater policy stringency—facilities must purchase more allowances on the carbon market when $R_{it} < 1$. The expected effect is that lower allocation ratios induce emissions reductions, leading to lower verified CO_2 and correspondingly lower satellite-derived NOx (as NOx is a combustion co-pollutant).

4.3.2 Discrete Treatment for Staggered DiD

For the Callaway-Sant'Anna estimator, I define a binary treatment indicator:

$$D_{it} = \mathbf{1}\{R_{it} < 1\} \tag{22}$$

where $\mathbf{1}\{\cdot\}$ is the indicator function. Treatment onset (cohort assignment) is defined as the first year a facility becomes treated:

$$G_i = \min\{t : D_{it} = 1\} \tag{23}$$

Facilities never treated are assigned $G_i = 0$ (the never-treated control group).

4.4 TWFE with Continuous Treatment

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

$$Y_{it} = \alpha_i + \gamma_{r(i),t} + \beta R_{it} + \mathbf{X}'_{it}\boldsymbol{\delta} + \varepsilon_{it} \tag{24}$$

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.

4.5 Callaway-Sant'Anna Difference-in-Differences

The Callaway and Sant'Anna [10] estimator addresses potential bias in standard TWFE when treatment timing varies across units and treatment effects are heterogeneous. The method estimates group-time average treatment effects:

$$ATT(g,t) = \mathbb{E}[Y_{it}(1) - Y_{it}(0) \mid G_i = g] \tag{25}$$

for each cohort g (first treatment year) and calendar time t.

Key features of the implementation:

- **Control group**: Never-treated units (facilities with $R_{it} \geq 1$ in all years)
- **Estimation method**: Doubly robust (combining outcome regression and propensity score weighting)
- **Anticipation**: Zero (no anticipation effects assumed)

The group-time ATTs are aggregated to produce:

1. **Simple aggregate ATT**: Overall average treatment effect across all treated observations:
$$ATT^{\text{simple}} = \sum_{g} \sum_{t \geq g} w_{g,t} \cdot ATT(g,t) \tag{26}$$
2. **Dynamic/event-study ATT**: Effects by time relative to treatment:
$$ATT(e) = \sum_{g} w_g \cdot ATT(g, g+e) \tag{27}$$

where e is event time (years since treatment onset)

The event-study specification enables pre-trend testing: under parallel trends, $ATT(e) = 0$ for $e < 0$. A Wald test of joint nullity for pre-treatment periods provides formal evidence on the parallel trends assumption.

4.6 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 ($\sim$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.

4.6.1 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 [25]. 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.

Let $r \in \{1, \ldots, R\}$ index NUTS2 regions. The cluster-robust variance estimator for the TWFE coefficient is:

$$\widehat{\text{Var}}(\hat{\beta}) = (\mathbf{X}'\mathbf{M}\mathbf{X})^{-1} \left(\sum_{r=1}^{R} \mathbf{X}_r' \hat{\mathbf{u}}_r \hat{\mathbf{u}}_r' \mathbf{X}_r \right) (\mathbf{X}'\mathbf{M}\mathbf{X})^{-1} \tag{28}$$

where $\mathbf{M}$ is the residual maker for fixed effects and $\hat{\mathbf{u}}_r$ are residuals for observations in NUTS2 region r.

4.7 Summary of Specifications

Table 2 summarizes the six core specifications estimated in this study. The dual-outcome design combined with two estimators and two embedding strategies for the satellite outcome yields six specifications.

Table 2 Summary of Econometric Specifications

Spec	Outcome	Embedding Strategy	Estimator
1	ETS CO_2	None	TWFE
2	ETS CO_2	None	Callaway-Sant'Anna
3	Satellite NOx	PCA (10 dims)	TWFE
4	Satellite NOx	PCA (10 dims)	Callaway-Sant'Anna
5	Satellite NOx	PLS (10 dims)	TWFE
6	Satellite NOx	PLS (10 dims)	Callaway-Sant'Anna

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. All specifications use Facility + Region×Year fixed effects and cluster standard errors by NUTS2 region.

5 Results

[Results to be added upon completion of empirical analysis.]
This section will present:

1. Descriptive statistics on treatment and outcome distributions
2. Estimates for all six specifications (Table 2)
3. Callaway-Sant'Anna event study results with pre-trend tests
4. Comparison of CO_2 and NOx responses: agreement in sign and magnitude
5. Stability of results across embedding strategies (PCA vs PLS)

5.1 Treatment Distribution

The allocation ratio exhibits substantial within-facility variation over the sample period. [Distribution statistics and figures to be added.]

5.2 Main Estimates

Table 3 reports the main estimation results.

5.3 Event Study

Figure **??** presents the Callaway-Sant'Anna event study results for both outcomes, plotting dynamic treatment effects by years relative to treatment onset. Comparison of CO_2 and NOx event-study patterns provides a cross-validation of policy effects: if

Table 3 Main Estimation Results: Effect of Allocation Ratio on Dual Outcomes

	Verified CO_2 ($ktCO_2/yr$)	Satellite NOx (kg/s)
Allocation Ratio (R_{it})	[–]	[–]
	([–])	([–])
Facility FE	Yes	Yes
Region×Year FE	Yes	Yes
Controls	Yes	Yes
Observations	[–]	[–]
Facilities	[–]	[–]
R^2 (within)	[–]	[–]

Note: Standard errors clustered by NUTS2 region in parentheses. Controls include capacity (MW), fuel shares (coal, gas), and AlphaEarth embeddings (64 dimensions). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

carbon pricing reduces combustion, both outcomes should show similar post-treatment declines.

[Event study figure to be added.]

Pre-trend tests: [Wald test results to be added.]

5.4 Robustness

[Robustness analysis to be added, including:]

- Alternative emissions thresholds (250 kt, 500 kt)
- Alternative treatment definitions (shortfall instead of ratio)
- Heterogeneity by fuel type and country
- Sensitivity to NOx detection limit (excluding facilities below 0.11 kg/s)
- Comparison of CO_2 vs NOx elasticities (percentage changes)

6 Discussion

[Discussion to be completed with empirical results.]

This study develops a methodological framework for evaluating climate policy impacts using dual outcomes: verified EU ETS CO_2 emissions and satellite-derived NOx emission proxies. This approach enables comprehensive assessment of both greenhouse gas reductions and air quality co-benefits. Several aspects merit discussion.

6.1 Interpretation of Estimates

The allocation ratio treatment variable has a natural interpretation in terms of policy stringency. A coefficient estimate of β on R_{it} implies that a 0.1 unit decrease in the allocation ratio (from 1.0 to 0.9, representing a shift from full free allocation to 10% shortfall) is associated with a -0.1β change in the outcome variable.

For verified CO_2: The outcome is directly measured and reported under EU ETS compliance, providing high-quality estimates of policy effects on greenhouse gas emissions.

For satellite-derived NOx: The Beirle-style flux-divergence approach provides a metric that is physically grounded in the continuity equation, with advection proportional to local emissions minus chemical sinks. Unlike simple column density comparisons, this method isolates facility-attributable emissions from regional background. However, interpretation requires caution due to the substantial uncertainty components discussed in Section 3.4.

Cross-validation: Agreement in sign and broad magnitude between CO_2 and NOx responses provides confidence that both outcomes are capturing genuine policy effects. Perfect correspondence is not expected—NOx and CO_2 have different emission factors by fuel type and combustion conditions—but directional agreement supports the validity of both measures.

6.2 Identification Concerns

Several potential threats to identification remain. First, **operational confounding**: facilities may respond to high carbon prices by adjusting operations (reducing output, switching fuels) in ways that are not fully captured by the control variables. 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.

Second, **measurement error in satellite NOx**: The Beirle-style flux-divergence approach is subject to multiple uncertainty sources (Section 3.4): lifetime parameterization (50% relative uncertainty), product differences between OFFL and PAL ($\pm$25%), NOx/NO_2 scaling ($\sim$7%), and spatial integration noise. These compound to total uncertainties of 20–40% per facility-year, with potential systematic low bias of $\sim$20% relative to reported emissions. This measurement error attenuates treatment effect estimates toward zero, making the satellite-based findings conservative.

Third, **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$\times$year fixed effects specification partially addresses this by absorbing regional substitution patterns.

6.3 Limitations of the Satellite NOx Proxy

The satellite-derived NOx estimates should be interpreted with several limitations in mind:

- **Use of OFFL L3 instead of PAL**: Beirle et al. (2023) use the PAL NO_2 product, which provides 10–40% higher TVCDs. The OFFL L3 product available via Google Earth Engine introduces additional uncertainty.
- **Simplified NOx/NO_2 scaling**: The photostationary-state approximation may not hold near stack plumes; I apply a fixed ratio of 1.38 $\pm$ 0.10 following Beirle et al. (approximately 7% uncertainty).
- **Integration radius trade-off**: The 15 km radius balances capturing the full point source signal against spatial interference from neighboring sources. For facilities in dense industrial areas, some signal contamination is likely.

- **Spatial interference**: For facilities with another ETS facility within 20 km (flagged via `interfered_20km`), the satellite outcome reflects emissions of the *local cluster* rather than an isolated point source. Treatment effect estimates for these facilities should be interpreted as cluster-level responses. Sensitivity analyses excluding interfered facilities test whether results are robust to this contamination.
- **Detection limit**: Beirle et al. Sect. 3.11.1 report detection limits of 0.11 kg/s (standard conditions) down to 0.03 kg/s (ideal high-albedo conditions). Since we do not implement a surface reflectivity mask, we flag observations against both thresholds and present sensitivity analysis using the conservative 0.11 kg/s threshold for Europe.
- **Skipping automatic identification**: We skip Beirle's automatic point-source identification algorithm because we have known ETS/LCP source locations. This design choice is explicitly endorsed by the authors and is conceptually appropriate for our use case, but it means we do not benefit from their classification step that filters artifacts and non-point sources. To compensate, we apply simplified significance filters (detection limit, statistical integration error, interference flag).
- **Unmodeled corrections**: AMF correction and plume height effects following Beirle et al. (2023) are not fully implemented, contributing to structural uncertainty. Topographic correction is implemented using SRTM elevation data.
- **Temporal averaging approach**: Beirle et al. compute temporal means of A* at orbit resolution before applying a single lifetime correction using mean wind speed. I instead aggregate to daily resolution and apply the lifetime correction per-day before temporal averaging, i.e., $E = \text{mean}[c_\tau(w_d) \times A^*_d]$ rather than $c_\tau(\bar{w}) \times \text{mean}(A^*)$. Since c_τ is nonlinear in wind speed, these are not mathematically identical; however, the difference is small for typical wind speed variance and the per-day approach is more appropriate for constructing annual panel estimates rather than a single multi-year catalog value.

These design choices—using OFFL instead of PAL, simplified significance flags, skipping the automatic identification algorithm—primarily increase noise and potential attenuation bias, not spurious detection. The satellite outcome remains a *physically grounded but noisy proxy*; ETS verified CO_2 remains the primary outcome for causal inference.

6.4 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 [8] demonstrates that pre-trained geospatial representations can serve as effective high-dimensional controls in panel settings. Unlike manually-specified geographic controls (distance to coast, elevation, population density), foundation model embeddings capture complex, nonlinear combinations of features that the model learned to be predictive across diverse tasks. The key assumption—that

these embeddings capture confounders affecting both policy exposure and air quality outcomes—is plausible given that the embeddings encode land use, infrastructure, and climate patterns that likely correlate with both industrial activity and pollution dispersion. Future work should investigate the sensitivity of causal estimates to different embedding specifications and 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 both fixed effects and inference represents a novel application of model-derived features for econometric analysis. Traditional approaches to clustered standard errors use geographic proximity or administrative boundaries that may not reflect the economically-relevant correlation structure. By using clusters derived from transmission network topology [9], which minimize within-cluster electrical distance, I define regions where facilities face correlated prices, dispatch patterns, and demand shocks—precisely the correlations that motivate cluster-robust inference [18]. This approach could be extended to other networked industries where external models of the network structure are available.

6.5 Limitations and Future Work

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 will have 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. The 500m spatial clustering threshold, while appropriate for grouping co-located emission sources into coherent facilities, may occasionally merge distinct plants or fail to group plants that share a common plume.

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

Satellite measurement uncertainty. The satellite NOx proxy is subject to the multiple uncertainty components detailed in Section 3.4, though these primarily introduce classical measurement error that attenuates rather than biases estimates.

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 ($\geq$2 m/s requirement), and satellite orbit patterns. This heterogeneity raises a potential methodological concern: if the subset of observable days differs systematically across facilities—e.g., coastal plants observed more frequently in winter, inland plants more in summer—then annual mean emission rates may not be directly comparable across locations. 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, so facility fixed effects absorb level differences; and (iii) year

fixed effects absorb common temporal patterns. As a robustness check, I include the number of valid observation days (`n_days_satellite`) as a control variable; stability of treatment effect estimates with and without this control supports the assumption that observation heterogeneity does not induce substantial bias.

UK exclusion. The EU ETS registry data used in this study does not include installations in the United Kingdom following Brexit. The UK withdrew from the EU ETS on January 1, 2021, and established its own UK Emissions Trading Scheme [26]. Consequently, UK large combustion plants—which represented a significant share of EU ETS-regulated capacity prior to 2021—are excluded from the analysis panel. Future work extending this framework to include UK facilities would require obtaining compliance data from the UK ETS registry and harmonizing it with the EU data sources used here. This may be facilitated in the near future by the linking of the UK and EU's ETS systems.

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, two of which follow a recent trend in causal inference toward incorporating machine learning-derived features to address high-dimensional confounding [5–7].

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 introduce network-based clustering derived from an external power system model for econometric inference. Standard errors are clustered using PyPSA-Eur power system clusters [9]—k-means clusters computed on transmission network topology features—that group facilities facing correlated wholesale prices, dispatch patterns, and grid constraints. This 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. 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 inference in policy evaluation.

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.

[Substantive empirical conclusions to be added with results.]

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 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 audited & corrected 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 [27, 28]. 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.

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.

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.

References

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

[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] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

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

[26] 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.

[27] 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.

[28] 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.