



Global variability in the detectability of power plant NO₂ plumes from space

Ruizhe Huang¹ and Sherrie Wang¹

¹Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Correspondence: Ruizhe Huang (rzhuang@mit.edu) and Sherrie Wang (sherwang@mit.edu)

Abstract. We present the first global, data-driven analysis of power plant NO₂ plume visibility from space. Using TROPOMI observations over 6,000 of the world's highest-emitting power plants and hourly CEMS data for 500 U.S. plants, we develop an automated algorithm that labels plumes and attributes them to their sources with 98% accuracy. We then train a machine learning model to predict plume detectability from environmental, meteorological, and observational variables (F1 score > 0.65, AUC > 0.8). Out of 25 variables, we find that NO_x emission rate, surface albedo, wind speed, and sensor zenith angle jointly explain much of the detection variability. An hourly NO_x emission rate of ≈400 kg/h corresponds to a 50% detection probability on average, but detection rates vary from < 20% to > 60% under different combinations of these conditions. These results provide the first empirical quantification of the physical and environmental factors that govern NO₂ plume visibility in satellite data, establishing a foundation for models to use similar predictors as auxiliary variables when quantifying emission rates from plume appearance.

1 Introduction

Fossil-fuel power plants are among the largest emitters of anthropogenic air pollution and greenhouse gases, responsible for approximately 22% of NO_x and over 40% of CO₂ emissions globally (McDuffie et al., 2020; M et al., 2022). Accurate quantification of these emissions is essential for enforcing environmental regulations and achieving international climate goals (Agreement, 2015; U.S. Environmental Protection Agency, 1970; European Commission, 2010). However, current global emission inventories exhibit substantial spatial and sectoral gaps, particularly in developing and rapidly industrializing regions (Cusworth et al., 2021; Guo et al., 2023; Couture et al., 2024). These gaps stem in part from the limitations of bottom-up accounting, which typically relies on fuel consumption statistics, emission factors, and engineering-based estimates (Guevara et al., 2024). In high-income countries, more robust systems such as Continuous Emission Monitoring Systems (CEMS)—which measure stack-level emissions in real time using instruments like NO_x analyzers, flow meters, and opacity monitors—enable direct, high-frequency reporting (Ding et al., 2025). In contrast, most power plants in low- and middle-income countries lack such infrastructure, leading to emissions inventories that are often uncertain, inconsistent, or outdated (Srivastava et al., 2024; Cusworth et al., 2021).

Satellite observations offer a promising top-down alternative for monitoring emissions at global scale. Recent studies have demonstrated the use of satellite instruments like TROPOMI and OMI to detect NO₂ enhancements over individual power



plants and estimate emissions through inverse modeling or statistical fits (Liu et al., 2020; Beirle et al., 2019; Goldberg et al., 2019). These works typically focus on a small number of high-emitting facilities under favorable meteorological and observational conditions, revealing both the potential and limitations of current detection capabilities (Rouet-Leduc and Hulbert, 2024; Cusworth et al., 2023; Dumont Le Brazidec et al., 2025).

30 Amidst the current satellite-based efforts to monitor anthropogenic NO_2 emissions, a challenge persists: power plant plumes with similar NO_x emissions exhibit vastly different detectability across regions (Figure 1), driven by meteorology, sensor geometry, and interference from nearby sources. While previous studies have demonstrated detection feasibility for individual plants, they have not systematically characterized the conditions that enable or inhibit detection across thousands of power plants globally. We address this scalability gap by asking: what meteorological, environmental, power plant, and sensor geometry variables determine whether a plume of given emission strength will be detectable? Answers to this question have
35 implications for improving satellite-based NO_x inverse modeling as well as future sensor design.

Specifically, we aim to systematically explain the spatiotemporal variability of NO_2 plume detectability from power plants using satellite observations. To do so, we develop a machine learning pipeline that first labels NO_2 plumes in TROPOMI (Sentinel-5P) images and then predicts plume detectability using features including meteorology, sensor geometry, environment
40 variables, and meteorology variables. We apply this framework to two datasets: (1) U.S. power plants with hourly CEMS data and (2) global high-emitting power plants using a 2018 annual NO_x inventory.

Our contributions are as follows.

1. We present a plume-labeling algorithm that achieves $\sim 98\%$ accuracy against manual annotations, designed for point source NO_2 attribution.
- 45 2. We produce the first global maps of plume detectability across 6,000 power plants, revealing strong geographic patterns. We find that an average hourly emission rate of ≈ 400 kg/h corresponds to a 50% detection probability on average, although detectability varies widely depending on location, sensing conditions, and meteorology.
3. We quantify how meteorological, environmental, and observational features influence NO_2 plume detectability across both regional and global scales. Using Permutation Score and SHAP Score Estimation, we identify NO_x Emission,
50 Surface Albedo, NO_2 Window, Sensor Zenith Angle, and Wind Speed as the dominant environmental and observational features that govern plume detectability.

By systematically characterizing where and when plumes are visible from space, this work clarifies the observational limits of current satellite NO_2 sensors. Furthermore, it identifies the environmental and geometric conditions under which power plant emissions are most likely to be accurately observed, providing empirical relationships that can inform future satellite-based
55 emission quantification.

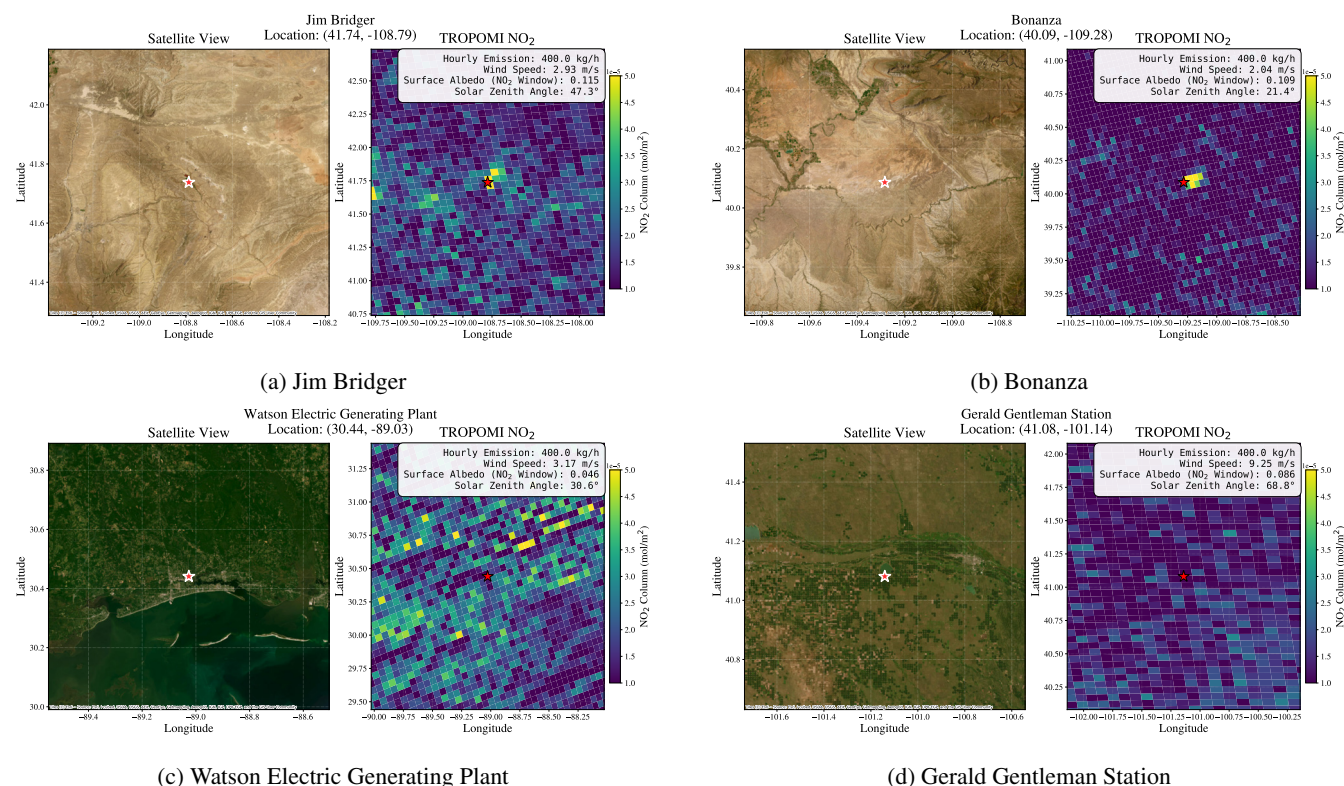


Figure 1. Comparison of tropospheric NO₂ vertical column densities observed by satellite over four U.S. power plants. Although all facilities reported the same hourly emissions (400 kg/h), NO₂ plumes are visible only near the Jim Bridger and Bonanza (top row). In contrast, no significant NO₂ enhancement is detected in the immediate area of the Watson Electric Generating Plant and Gerald Gentleman Station (bottom row). We seek to explain such differences in plume detectability. Satellite Image Source: Esri, i-cubed, USDA, USGS, AEX, GeoEye, Getmapping, Aerogrid, IGN, IGP, UPR-EGP, and the GIS User Community | Powered by Esri.

2 Datasets

2.1 U.S. and Global Power Plants

For our U.S. analysis, we tracked hourly NO_x emissions for the 500 power plants with the largest total NO_x emissions from 2019 to 2024, a group representing the upper tail of all U.S. facilities (Figure 2(c)). By 2024, pipeline natural gas facilities (305 plants, 61%) and coal plants (158 plants, 32%) dominated this cohort, with a variety of other fuels used by the few remaining plants such as natural gas (14), coal refuse (4), diesel oil (2), and other gas (2). Over this six-year period, however, this group of top emitters transitioned measurably from coal to natural gas. Within the cohort, coal-fired plants fell from 180 to 158 (a 12.2% decrease), while pipeline-connected natural gas facilities grew from 286 to 305 (a 6.6% increase).

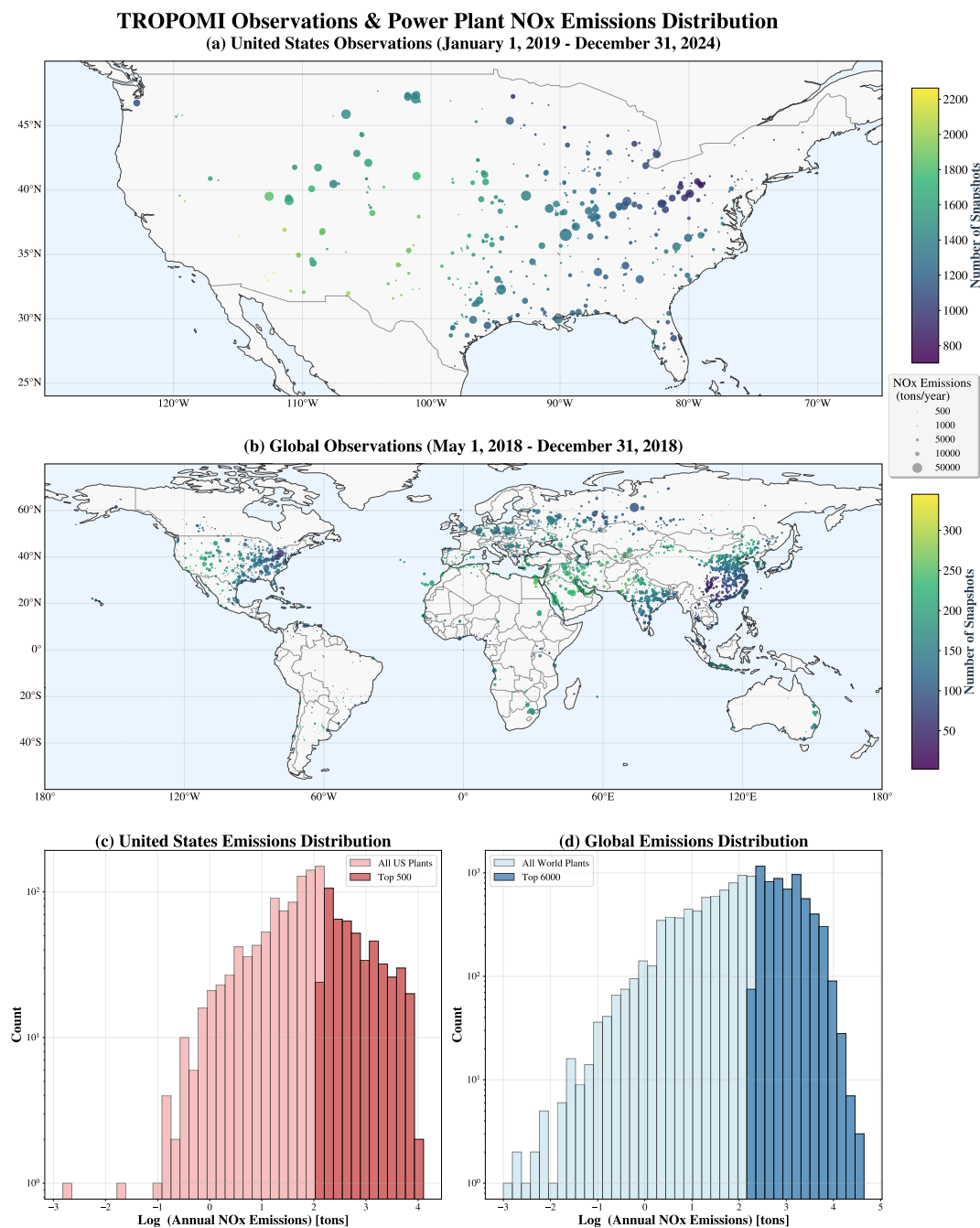


Figure 2. Distribution of TROPOMI observations and power plant NO_x emissions. Top panels show spatial distribution of power plant NO_x emissions (marker size) overlaid with TROPOMI observation frequency (marker color) for **(a)** United States and **(b)** global facilities. Bottom panels display emission distribution histograms on a log scale showing all plants (light colors) versus analysis subsets of top emitters (dark colors) for **(c)** United States (top 500 plants) and **(d)** global (top 6000 plants). TROPOMI observation periods: US data from January 2019 to December 2024, global data from May 2018 to December 2018.



For our global analysis, we rely on annual emissions data due to differences in data availability compared to the United States. We used the CoCO₂ power-plant emissions catalogue to identify the 6,000 facilities with the largest annual NO_x emissions in 2018 (Guevara et al., 2024). This group of plants, with emissions ranging from 219 to 44,548 tons/year, constitutes the upper tail of global power plant emissions (Figure 2(d)). Coal (3,023), natural gas (1,810), and oil (690) plants dominate the cohort, which also includes facilities that burn biomass (354) and waste (123).

To ensure data quality, we filtered this list for duplicates through a two-step process: we first identified plants within 3 km of each other, then removed any in those groups that shared identical emission values across all reported pollutants (CO₂, NO_x, SO_x, CO, and CH₄). This filtering removed 3,320 power plants from the original 16,461 facilities (20% of the catalog). China experienced the largest impact with 1,696 plants filtered out (45% of its facilities), followed by India with 498 removals (48%), Russia with 163 (25%), Japan with 160 (13%), and Indonesia with 101 (39%). Together, China and India accounted for 66% of all removed plants (2,194 out of 3,320), highlighting significant data quality issues concentrated in these two countries. In contrast, the United States had minimal filtering with only 1 plant removed out of 2,850, suggesting better data quality in the U.S. portion of the COCO₂ catalog.

2.2 Satellite NO₂ Measurements

To form the observation dataset, we used satellite data of tropospheric NO₂ vertical column densities from the TROPOMI instrument aboard the Sentinel-5P satellite. Launched in October 2017, TROPOMI provides data at a spatial resolution of 7 km × 3.5 km, which improved to 5.5 km × 3.5 km in August 2019.

We downloaded raw Level-2 NO₂ data in NetCDF format from the NASA Earthdata portal. Each file corresponds to a single satellite orbit (~101 minutes), containing one Level-2 swath with 450 across-track ground pixels and thousands of along-track scanlines. For our U.S. analysis (January 1, 2019 – December 31, 2024), we acquired 9,636 files. For the global analysis (May 1 – December 31, 2018), we obtained 3,421 files.

To generate a high-quality dataset, we processed these raw files through a multi-step filtering pipeline. For each TROPOMI observation over a power plant, the observation is kept if it contains:

1. a quality flag greater than 0.75, which removes data contaminated by clouds or other errors,
2. at least 50% valid pixels within a 50 km radius of a power plant, and
3. a valid nearest-neighbor pixel over a power plant.

After applying these quality assurance criteria, we dropped any remaining observations containing null (NaN) values in extracted features. This entire process yielded a final dataset of 504,165 high-quality observations for the U.S. analysis and 875,580 for the global analysis. Across the mapped facilities (U.S., $N = 500$; global, $N = 6,000$), snapshot coverage is extensive. In the U.S., plants have a median of 1,301 snapshots ($p_{10} = 1,050.0$, $p_{90} = 1,644.7$); every plant has ≥ 700 snapshots (range 701–2,263). Globally, the median is 141 snapshots ($p_{10} = 90$, $p_{90} = 218$), with 98.8% of plants having ≥ 50 snapshots



95 (range 2–345). As illustrated in Figures 2(a) and 2(b), the geographic density of these final observations is highest in drier, less cloudy regions and at higher latitudes.

2.3 Hourly NO_x Emissions in the U.S.

The U.S. Environmental Protection Agency’s (EPA) Clean Air Markets Program Data (CAMPD) provides hourly emissions data for CO_2 , NO_x , SO_2 , and mercury from individual power plants. This level of granularity is unique globally and offers
100 a distinct opportunity to study how emission rates influence the detectability of NO_x plumes in satellite imagery. For our analysis, we downloaded hourly NO_x emission data from the CAMPD API, specifically matching the times of TROPOMI satellite observations for each plant.

Continuous Emissions Monitoring Systems (CEMS) provide these hourly emissions data by measuring pollutant concentrations and exhaust flow rates to compute mass emission rates. These systems operate continuously, reporting hourly averaged
105 data directly to the EPA.

The CAMPD inventory is comprehensive, containing 97.1% of annual U.S. NO_x emissions as of 2023 (U.S. Environmental Protection Agency, 2023). From this inventory, we selected 500 power plants for our study. Our selection accounts for 93.6% of the inventory’s NO_x emissions and 79.8% of its CO_2 emissions. For context, the EPA’s broader power sector data covered approximately 96% of U.S. fossil fuel-based electricity generation as of 2018 (epa).

110 We began our analysis with 641,899 TROPOMI overpasses for our selected power plants. After removing instances with incomplete data (i.e., any NaN values), 641,424 overpasses remained. We then successfully matched 504,165 of these to corresponding hourly records in the CAMPD database. We could not match the remaining overpasses because CAMPD only catalogues regulated facilities during operating hours with valid CEMS data; overpasses during monitoring downtime, plant shutdowns, or over facilities below reporting thresholds lacked corresponding entries. Our final U.S. dataset therefore contains
115 504,165 samples, each pairing a TROPOMI image with that hour’s reported NO_x emissions.

2.4 Annual NO_x Emissions for Global Power Plants

Because global hourly emissions are unavailable, we used annual emissions. For our global analysis, we sourced annual power plant emissions from the CoCO₂ catalog (Guevara et al., 2024). This comprehensive inventory covers at least 95.9% of total global power plant emissions and provides annual data for CO_2 , NO_x , SO_x , CO, and CH_4 . The CoCO₂ catalog derived the
120 annual emissions using a bottom-up methodology as follows:

1. **Europe:** Use each plant’s official reported NO_x from EPRTTR and LCP dataset (European Environment Agency, 2020, 2019) for the year. If a plant didn’t report, estimate it from other official datasets or standard country-and-fuel ratios.
2. **United States:** Use the EPA’s eGRID (largely CEMS-based) (U.S. Environmental Protection Agency, 2020).
3. **Rest of world:** Start from national fuel use, convert to NO_x with country- and fuel-specific ratios, then apportion to
125 plants by installed capacity.



These derived annual emissions provide robust annual magnitudes but do not resolve within-year variability. The plume detection is near-instantaneous; if we only have annual emissions, this will limit our ability to predict the plume detection.

From this catalog, we selected 6,000 plants that account for 96.1% of the inventory's NO_x and 91.5% of its CO_2 emissions. This selection represents a significant portion of worldwide power plant pollution, corresponding to at least 92.1% of global NO_x and 87.7% of global CO_2 emissions. We then gathered 875,683 TROPOMI overpasses for these plants and removed a small number with incomplete data. Our final global dataset contains 875,580 samples, each pairing a TROPOMI image with the plant's annual NO_x emissions.

2.5 Features Predictive of Plume Detectability

We selected features to predict plume detectability based on the physical processes that govern plume formation, transport, and satellite detection (Table 1). These features fall into four categories: sensor characteristics, power plant attributes, meteorological conditions, and environmental context. The variables in these categories capture the complex interplay between emission sources, atmospheric conditions, and observation geometry that strongly influences plume detectability.

We extracted these variables from their respective data sources using several methods:

1. TROPOMI features were extracted from the pixel closest to each power plant's center.
2. ERA5 meteorological fields were interpolated to plant locations using nearest-grid interpolation.
3. Emission data from EPA or global inventories were used directly.

We describe the variables and their source datasets in Table 1, organized by category.

2.5.1 Sensor

Sensor zenith and azimuth angles define the viewing geometry and optical path length. Along with solar angles and scene albedo, they control the amount of backscattered light available for absorption spectroscopy. Sensor altitude affects spatial resolution and field of view, while the scaled small-pixel variance measures sub-pixel heterogeneity that can indicate a plume's presence. Together, these variables establish the instrument's sensitivity to narrow, high-contrast plumes.

2.5.2 Power plant

We include the primary fuel type and emission rate for each power plant. For U.S. facilities, we use hourly emissions; for global facilities, we use annual emissions. The primary fuel type (e.g., coal, natural gas) influences combustion temperature and stack gas buoyancy, which affect plume rise and dispersion. Because higher emissions produce more concentrated and easily detected plumes, emission magnitude is a key predictor of detection success.



Table 1. Input features for NO₂ retrieval by category (Sensor, Power Plant, Meteorology, Environment) and source (TROPOMI, ERA5, U.S. EPA, Global Inventory, and SimpleMaps).

Category	Source	Features
Sensor	TROPOMI	sensor azimuth angle, sensor zenith angle, sensor altitude, scaled small-pixel variance
Power plant	U.S. EPA ^a	hourly NO _x emissions, primary fuel type
	Global Inventory ^b	annual NO _x emissions, primary fuel type
Meteorology	TROPOMI	10 m wind speed, cloud albedo ^c , cloud pressure ^c , cloud fraction ^c , surface pressure, solar zenith angle, solar azimuth angle, apparent scene pressure, aerosol index 354–388
	ERA5	2 m temperature, total column water vapour, TOA incident solar radiation
Environment	TROPOMI	surface classification, snow/ice flag, scene albedo, surface albedo, surface albedo nitrogen-dioxide window, surface altitude, surface altitude precision

^a Data source for the U.S. analysis

^b Data source for the global analysis (Guevara et al., 2024)

^c Input variables for Clouds-as-Reflecting-Boundaries (CRB) model, a simplified Lambertian-reflector approach used by the ROCINN algorithm to retrieve cloud properties

2.5.3 Meteorology

We include meteorological variables to account for their influence on plume transport, optical interference, and radiative transfer.

- **Transport and Mixing:** Wind speed drives horizontal advection and dilution. Temperature influences boundary layer dynamics and vertical mixing, which together determine a plume’s rise and spread. We use 10 m horizontal winds (u , v) available from the TROPOMI product.
- **Optical Interference:** Cloud properties (fraction, pressure, and albedo) can obscure plumes or create false signals. The aerosol index helps identify absorbing particles that interfere with spectroscopic measurements.



- **Radiative Transfer:** Solar geometry (zenith and azimuth angles) determines the photon path length, while scene pressure and surface pressure define the effective reflecting altitude and total depth of the atmospheric column. Total column water vapor affects NO₂ spectral signatures, while top-of-atmosphere incident solar radiation quantifies the energy available for the measurement.

165 2.5.4 Environment

We include environmental variables to characterize surface properties.

- **Surface Properties:** Surface characteristics affect the accuracy of NO₂ retrieval. We use surface classification and snow/ice flags to account for reflectance patterns and potential spectral interferences. We also use surface altitude to establish the vertical column depth. Finally, multiple albedo measurements quantify surface reflectivity, which controls the signal strength for detecting a plume against different backgrounds.

170

3 Methods

In this section, we describe our methodology for quantifying the factors that control the detectability of NO₂ plumes in satellite observations. Our systematic approach determines when power plant emissions can be detected from space, and these results help understand the capabilities and limitations of current satellite monitoring.

175 As illustrated in Figure 3, our analytical workflow consists of four main stages. First, we filter TROPOMI NO₂ observations and extract meteorological and environmental features. Second, we apply our Automated Plume Labeling Algorithm to identify detectable plumes in each satellite overpass. Third, we train machine learning models to predict plume detectability from extracted features. Finally, we use the trained models to quantify the relative importance of each feature.

We implement this pipeline in two complementary analyses:

180 **U.S. Analysis:** Uses hourly NO_x emission data (2019–2024) from 500 high-emitting power plants with Continuous Emissions Monitoring Systems (CEMS).

Global Analysis: Uses annual NO_x emission inventories (2018) for 6,000 high-emitting power plants worldwide.

3.1 Automated Plume Detection

To build a machine learning model to explain plume detectability, we first need a large labeled dataset of plume detection. Manual plume annotation is prohibitively costly and inconsistent, so labels must be generated automatically. Existing automated methods such as DDEQ (Kuhlmann et al., 2024) detect NO₂ anomalies but cannot attribute a plume to a specific source, failing to separate a target power plant from nearby cities or other industries. To fill this attribution gap, we develop a new plume-detectability-with-attribution algorithm that automatically labels plumes and assigns each one to its originating power plant.

185

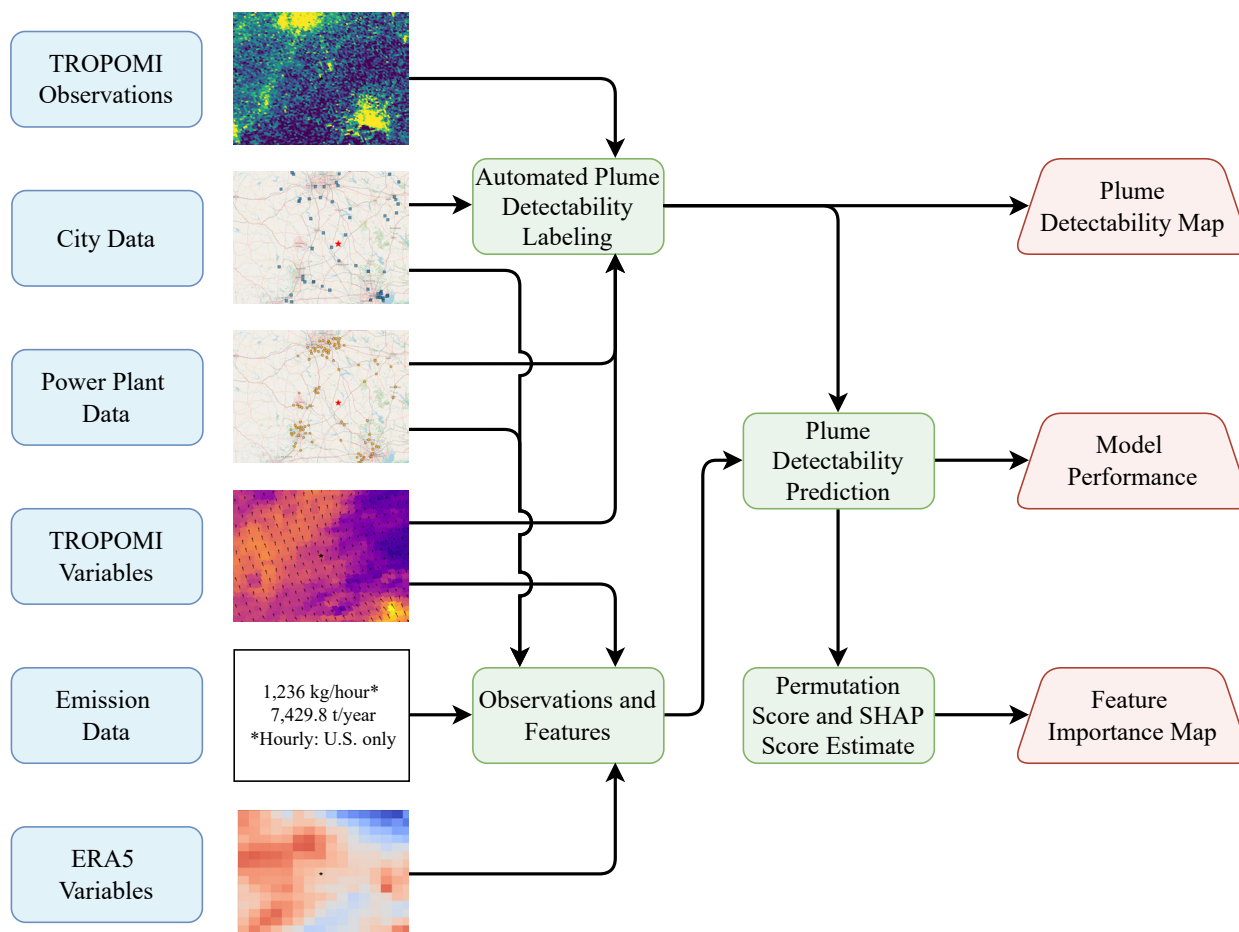


Figure 3. Overview of the analytical pipeline. The process uses input data (blue) to feed the main pipeline processes (orange). First, the Automated Plume Labeling algorithm produces detectability labels for each observation. These labels, combined with extracted features, are then used to train a predictive model (Plume Detectability Prediction). In a final step, this trained model is analyzed using permutation and SHAP scores to estimate the importance of each feature (Permutation Score and SHAP Score Estimate). The final products (red) originate from these distinct stages: the labeling algorithm generates a plume detectability map, the predictive model's outputs are used for a performance evaluation, and the final analysis generates the quantification of feature importance.



190 Our algorithm automates this attribution through a five-step workflow, outlined in Figure 4. The process integrates TROPOMI satellite measurements with power plant and city databases to systematically isolate and label a target plume. While the steps are executed sequentially, the figure illustrates the key data dependencies that connect the different stages of the process.

3.1.1 Hyper-parameter Tuning

To tune the hyper-parameters of the Automated Plume Detectability Algorithm, we drew a stratified sample of 100 snapshots
195 from the global dataset—balanced across six continents and five emission-level quantiles by selecting roughly 3-4 snapshots per stratum—and an additional 100 from the U.S. dataset. Stratified sampling here means dividing the dataset into strata defined by continent and emission quantile, then sampling equally from each to ensure representative coverage. Hyper-parameters were then selected using heuristic physical approximations and insights from this tuning set.

3.1.2 Step 1: Masking Interference Zones

200 To isolate the target power plant’s signal, the algorithm first identifies and masks potential sources of interference. It begins by cataloging all major cities and other power plants within a **150 km** radius of the target facility.

The algorithm flags a source as interfering based on specific criteria. It considers a city an interference source if its population exceeds **200,000**, a threshold set to capture major urban areas whose NO₂ emissions from traffic and industry could be mistaken for the plant’s plume. Similarly, it flags another power plant as an interference source if its emission rate is equal to or greater
205 than the target plant’s (an emission ratio of **1.0** or higher). This step prevents the algorithm from misattributing signals from stronger nearby emitters to the target facility. We sourced city and corresponding population data from the SimpleMaps World Cities Database (SimpleMaps, n.d.), a global dataset of approximately 48,000 cities last updated on May 19, 2025. The full power plant inventories are taken from the U.S. EPA for U.S. analysis and the CoCO₂ catalog (Guevara et al., 2024) for global analysis, mentioned in Section 3.

210 To exclude these confounding signals, the algorithm applies a distinct spatial mask to each source type:

- **City Masking:** The algorithm applies an adaptive mask to cities, where the radius scales with the city’s log-transformed population to account for the larger pollution footprint of major urban centers. It calculates the radius to be between **10 km** and **90 km** by applying a scaling factor of **9.0** from a **zero** baseline to the log population. This method ensures larger metropolitan areas are masked with a wider radius.
- 215 – **Power Plant Masking:** For other interfering power plants, the algorithm applies a fixed-radius mask of **20 km**. This standard distance excludes the immediate near-field zone where plumes from separate plants are most likely to mix and become indistinguishable.

This adaptive masking ensures that the subsequent analysis—conducted within a **100 km** window and characterized at **50 km**—excludes pixels influenced by upwind pollution sources unrelated to the target plant. Examples of city and power plant
220 masking are shown in Appendix F.

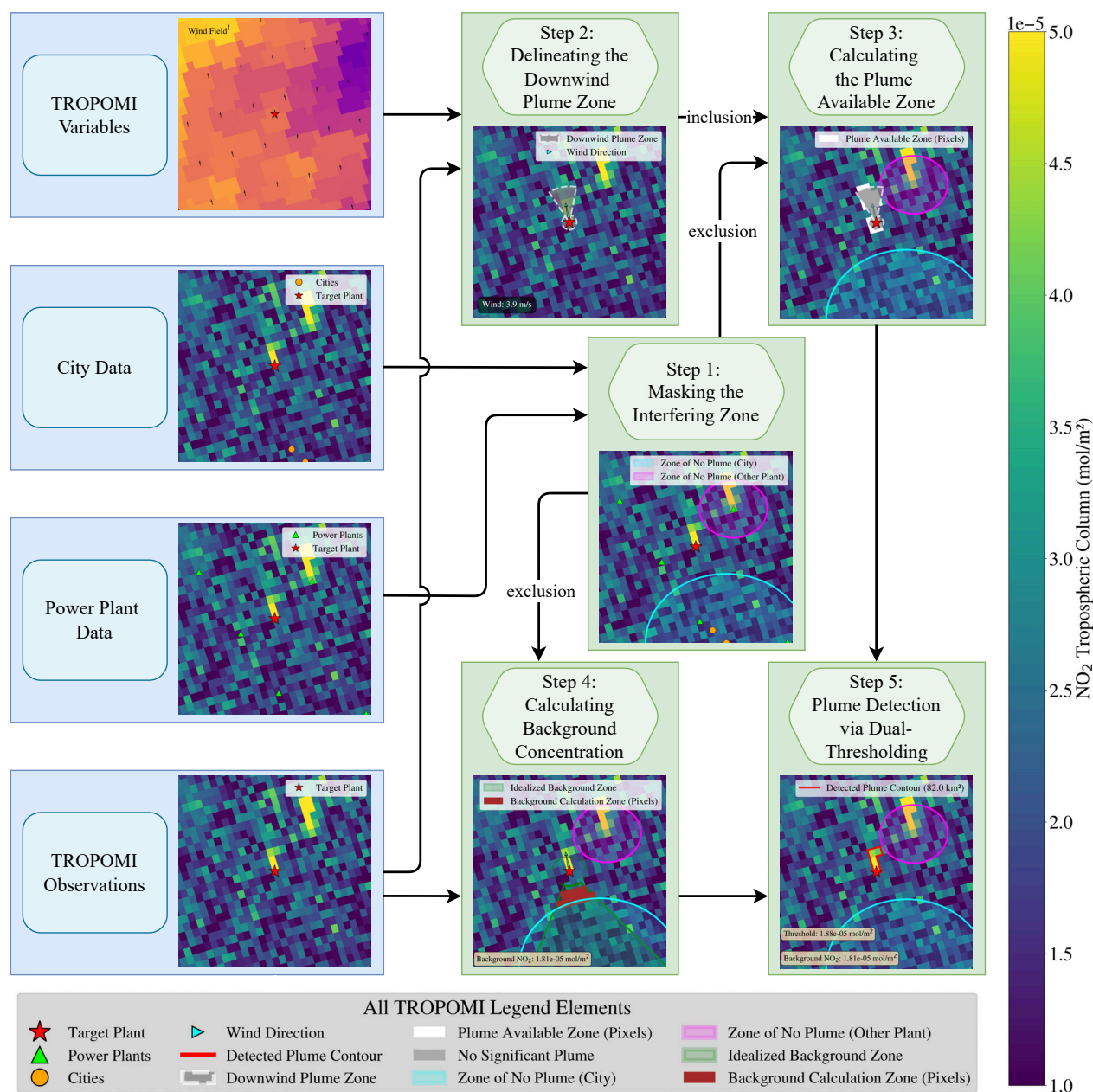


Figure 4. Overview of the Automated Plume Detection Algorithm. The process begins by masking potential interference from nearby cities and power plants (Steps 1). It then uses wind direction to define a downwind search area and isolates a “Plume Available Zone” by removing the masked regions (Steps 2-3). Finally, the algorithm applies a dual-thresholding method against the calculated background concentration within the Background Calculation Zone to detect and label the plume (Step 4-5). Blue boxes represent input data, and green boxes represent algorithm components. A full legend and color bar are provided.



3.1.3 Step 2: Delineating the Downwind Plume Zone

In this step, the algorithm uses meteorological data to define the geographic area where the target plant's plume is expected to travel. It retrieves local wind vectors (speed and direction) directly from the TROPOMI satellite product for the time of the overpass.

225 Using this wind direction, the algorithm establishes a downwind search sector. It applies a tolerance of $\pm 25^\circ$ to the wind vector, creating a 50° angular cone. The algorithm discards any potential signals falling outside this cone as inconsistent with the plume's likely trajectory. Within this sector, the search for a plume is limited to a maximum distance of **20 km** from the source, with special consideration given to close-range transport within **5 km**. Finally, any detected enhancement must cover a minimum area of **25 km²** to be flagged as a valid signal.

230 3.1.4 Step 3: Calculating the Plume Available Zone

The **Plume Available Zone** is the final, refined area where the algorithm searches for the target's plume. The algorithm defines this zone by taking the **Downwind Plume Zone** from Step 2 and subtracting the **Zone of No Plumes** from Step 1. This step restricts the analysis to only those pixels that are both meteorologically downwind of the target and spatially removed from confounding emission sources.

235 3.1.5 Step 4: Calculating Background Concentration

To accurately quantify a plume's enhancement, the algorithm first establishes the ambient NO₂ concentration in the upwind region. It calculates this background value by sampling pixels from a **120°** upwind sector located in an annulus between **10 km** and **100 km** from the plant. The algorithm defines this upwind sector as the direction opposite the plume's flow ($\pm 60^\circ$ tolerance).

240 To prevent contamination, the algorithm excludes any pixels within this upwind sector that were masked as interference zones in Step 1. It then defines the background concentration as the median of the remaining filtered pixels. Using the median ensures the background value is robust against outliers, providing a reliable baseline.

3.1.6 Step 5: Plume Detection via Dual-Thresholding

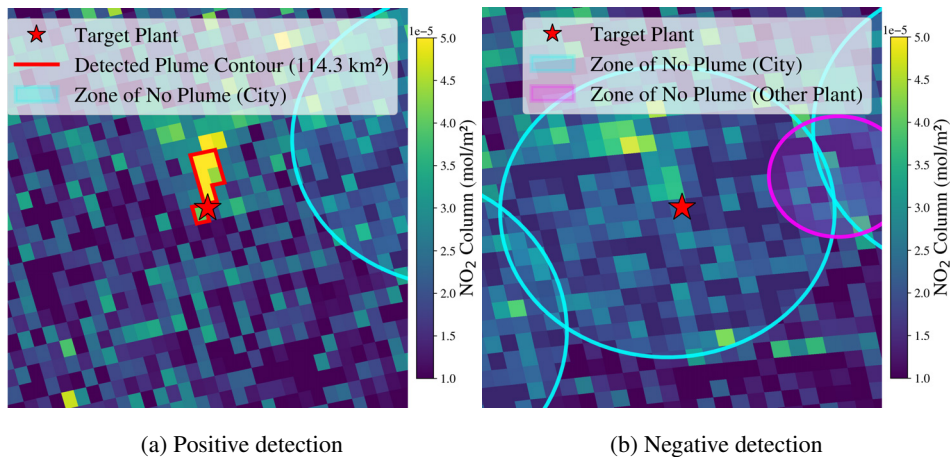
In the final step, the algorithm identifies the plume by applying a dual-criteria threshold to each pixel within the Plume Available
245 Zone. The algorithm flags a pixel as part of the plume only if its NO₂ concentration meets **both** of the following conditions:

1. **Statistical Significance:** The pixel's value must be at least **2 standard deviations (2σ)** above the local background concentration calculated in Step 4. This ensures the signal is statistically significant relative to normal atmospheric variability.
2. **Absolute Minimum:** The pixel's value must also exceed an absolute floor of **5×10^{-6} mol/m²**. This threshold pre-
250 vents false positives that could be caused by either minor fluctuations in very clean background air or by instrument



noise. We set this value at approximately twice the TROPOMI instrument’s average noise floor (a "stripe amplitude" of $2.15 \times 10^{-6} \text{ mol/m}^2$), as characterized by (Van Geffen et al., 2020), to robustly distinguish real signals from sensor artifacts.

By requiring a signal to be both statistically significant relative to the background and above a fixed minimum, this dual-
255 threshold method robustly identifies the emission plume while minimizing noise-based false detections.



(c) Validation on Random Sample (n=400). Results: 98.25% Accuracy, 96.67% Precision, 82.86% Recall.

	Pred. False	Pred. True	Total
Actual False	364	1	365
Actual True	6	29	35
Total	370	30	400

(d) Validation on Stratified Sample (n=400). Results: 97.75% Accuracy, 89.58% Precision, 91.49% Recall.

	Pred. False	Pred. True	Total
Actual False	348	5	353
Actual True	4	43	47
Total	352	48	400

Figure 5. Validation of the Automated Plume Detection Algorithm. (a) An example of a positive detection, where the algorithm correctly identifies a plume from a target power plant after masking interference. (b) An example of a negative detection, where no plume is attributed to the target due to interfering sources. (c) Confusion matrix for the validation on 400 randomly sampled records. (d) Confusion matrix for the validation on 400 records selected via stratified sampling to ensure geographic and emission-level balance.



3.1.7 Validation of the Plume Labeling Algorithm via Manual Labeling

To validate the plume-labeling algorithm, we conducted two tests on manually labeled datasets that were separate from the one used for hyperparameter tuning. Each test used a different sampling strategy to answer a specific question, one focused on performance under real-world data distributions, the other on robustness across a stratified sample balanced by region and emission level.

In the first test, we evaluated algorithm performance under realistic, operational conditions using 400 samples randomly drawn from the full test set of 175,116 records. This sampling preserved the natural distribution of power plants, resulting in a geographically and quantitatively imbalanced dataset. For example, the sample was skewed toward Asia (235 records) compared to North America (59) and Europe (57), with limited representation from other regions. Emission categories were similarly uneven, ranging from 93 samples in the lowest quantile (E1, 0–20%) to 66 in the next (E2, 20–40%). Manual verification of this set provides a benchmark for real-world performance. The resulting confusion matrix (Figure 5c) shows an overall accuracy of 98.25%.

In the second test, we assessed robustness by constructing a balanced validation set of 400 samples through stratified sampling. From the full test set, we drew roughly equal numbers of samples (13–14) from strata defined by seven continents and five emission-level quantiles (0–20%, 20–40%, etc.). This yielded a dataset with 80 samples in each emission bin, ensuring no single region or emission category dominated. On this balanced set, the algorithm achieved an accuracy of 97.75% (Figure 5d).

3.2 Plume Detectability Prediction

In this section, we evaluate how effectively our input features distinguish between detectable and non-detectable plumes (plume detectability). To do so, we conduct an **item split experiment**, training a model on a random sample of observations from all power plants.

We frame plume detectability as a binary classification task. For each TROPOMI observation, which our algorithm labels as either “plume detected” or “no plume,” we extract a comprehensive set of meteorological, instrumental, environmental, and emission-related features to use as classifier inputs.

We selected a Multi-Layer Perceptron (MLP) after finding it outperformed random forests and logistic regression. The MLP is a fully-connected feedforward network with four hidden layers (256, 128, 64, and 32 neurons) using ReLU activation functions. To mitigate overfitting, we apply dropout ($p = 0.3$) after the first three hidden layers. The network takes a 25-dimensional feature vector as input and produces a binary classification output. With a 25-dimensional input, the model has 53,761 learnable parameters. This architecture is used for all U.S. and global analyses.

To understand how performance varies with emission magnitude, we trained separate models on four datasets: “All” (our complete power plant inventory) and three subsets of the highest emitters (“Top-100”, “Top-50”, and “Top-20”). For each dataset, we partitioned observations into training (60%), validation (20%), and test (20%) sets. To isolate environmental effects and avoid confounding from nearby emission sources, within each dataset, we exclude plants whose centroids fall within predefined interference zones (cities and co-located plants; see Section 3.1). This conservative filter reduces misattribution,



ensuring detected plumes are attributable to the target plant. In this way, the environmental features are the dominant factors in
290 the plume detectability instead of the influence from nearby cities and other power plants.

We addressed class imbalance in the training set using random oversampling and then standardized all input features to have zero mean and unit variance, based on the resampled training data statistics. We trained each model to maximize the Area Under the Curve (AUC) on the validation set. To account for training stochasticity, we repeated each experiment five times and reported the mean and standard deviation of the results. Full training details are in Appendix B.

295 To further test our model, we assessed cross-plant transferability of plume detection features. We accomplished this by conducting a **power plant split** experiment that trains on one set of plants and tests on a separate set. In this experiment, we sought to understand how well a model trained on one set of power plants performs on another set of power plants not included in the training set.

In the main text, we focus on the item split experiment results for both U.S. and global snapshots (Table 3). The results
300 for the power plant split experiment are presented in Appendix D (Table D1). The tables organize performance by region (U.S. vs. Global), emission data resolution (annual vs. hourly NO_x), and dataset scope (All vs. Top-X emitters). We report six classification metrics, highlighting the three most critical for operational monitoring: Precision, Recall, and the F1 score. These three metrics directly assess the system’s real-world effectiveness by quantifying the rate of false positives (Precision), the ability to capture all true events (Recall), and the overall balance between these competing objectives (F1 score).

305 3.3 Permutation Score and SHAP Score Estimate

3.3.1 Model-wide Feature Importance via Permutation

To assess the model’s overall reliance on each input feature, we use permutation importance—a technique well-suited for complex models like MLPs. For a given feature x_i and a held-out validation set $\mathcal{D}_{\text{val}}$, we first calculate the model’s baseline performance, $\mathcal{M}_{\text{orig}}$, using a chosen metric like AUC. We then randomly shuffle the values of only feature x_i across all
310 examples in $\mathcal{D}_{\text{val}}$ to break its relationship with the target variable, and we re-evaluate the model to get a permuted performance score, $\mathcal{M}_{\text{perm}(i)}$. The permutation importance of feature x_i is the resulting drop in performance:

$$\Delta_i = \mathcal{M}_{\text{orig}} - \mathcal{M}_{\text{perm}(i)}. \quad (1)$$

A larger value of Δ_i signifies that the model depends more heavily on that feature for its predictions.

3.3.2 Plant-Level Feature Importance via SHAP

315 While permutation importance gives a global view, we use SHAP (SHapley Additive exPlanations) to explain predictions for individual power plants. SHAP assigns an importance value, ϕ_i , to each feature based on its marginal contribution to a specific prediction. For a plant with feature vector $\mathbf{x}$, these SHAP values explain how the model’s prediction, $f(\mathbf{x})$, deviates from the



baseline prediction, $E[f(\mathbf{X})]$ (the average prediction across all data):

$$f(\mathbf{x}) - E[f(\mathbf{X})] = \sum_{i=1}^d \phi_i. \quad (2)$$

320 The exact computation of these Shapley values requires re-evaluating the model on all possible subsets of features, which is computationally prohibitive for any non-trivial number of features. We therefore estimate them using the SHAP DeepExplainer algorithm. This method, based on a deep learning attribution technique called DeepLIFT, efficiently approximates SHAP values. It works by propagating the difference between the model's final output and a baseline reference output backwards through the network layers. This process effectively assigns contribution scores to each neuron and, ultimately, to each input feature, providing a fast and accurate approximation for neural networks like our MLP.

Each value ϕ_i quantifies how much feature i pushes the prediction away from this baseline. A positive ϕ_i increases the predicted probability of detection, while a negative value decreases it. Ranking the absolute SHAP values, $|\phi_i|$, reveals which features most strongly drive the outcome for a single plant.

330 To visualize these local explanations on a global scale, we aggregate the plant-level SHAP results onto a grid. First, we assign each power plant to a grid cell. Within each cell, we identify the most influential feature by finding which feature most frequently has the highest absolute SHAP value ($|\phi_i|$) across all plants in that cell. Next, we determine that feature's dominant directional impact (positive or negative) within the cell. The final map visualizes these results, with each cell colored by its dominant feature and marked with an arrow indicating its typical influence on plume detectability.

4 Results

335 4.1 Most power plants are near other emission sources

First, we find that for the U.S. top-500 NO_x emitters (2019–2024), 312/500 plants (62.4%) were flagged as in interference zones in 2019, declining to 283 by 2024 as the active cohort shrank from 500 to 461. Metropolitan proximity remains the primary driver of interference. Across six years, only 171 plants were never within interference zones, accounting for 49.1% of emissions from the original 500.

340 Interference is even more pervasive in the global dataset of 6,000 major power plants. In total, 4,935 plants (82.3%) lie within at least one interference zone; proximity to large urban centers dominates (4,491 plants), while co-located power plants also contribute substantially (2,719 plants). Notably, 2,275 plants are simultaneously affected by both cities and other plants, underscoring the difficulty of identifying interference-free monitoring conditions at scale. In 2018, only 1,065 plants remained outside interference zones, retaining just 21.1% of the original emissions.

345 These contrasting retention rates—49.1% for U.S. plants versus 21.1% for global plants—reflect the higher spatial clustering of emission sources and urban areas in the global dataset.

At a snapshot level, across the global dataset, interference screening removed 714,462 of 875,580 TROPOMI snapshots (81.8%), retaining 161,118. For the U.S. dataset, 305,827 of 504,165 snapshots (60.7%) were removed across 2019–2024.



Cohort-level aggregates for smaller U.S. subsets show similarly large removals (Table 2); per-year details for the top-500 are given in Table A1.

While this filtering is stringent, it is necessary to ensure interpretability and causal attribution. Our goal is to analyze how meteorology, emission rate, and environmental factors affect plume visibility — not to model column enhancements that may arise from overlapping sources. Without isolating interference-free scenes, plume detections could be driven by emissions from neighboring plants or urban NO₂ backgrounds, obscuring the true relationships between plant-level conditions and observed plume visibility. Thus, the reduced sample size reflects a trade-off: a smaller but cleaner dataset that supports robust inference about the mechanisms underlying visible plumes in satellite imagery.

Table 2. Cohort-level snapshot filtering (aggregated, 2019–2024).

Cohort	Unfiltered	Retained	Removed	Removal Rate
All (Global)	875,580	161,118	714,462	81.6%
Top 100 (Global)	17,456	6,110	11,346	65.0%
Top 50 (Global)	8,669	3,206	5,463	63.0%
Top 20 (Global)	3,220	797	2,503	77.7%
All (U.S.)	504,165	198,338	305,827	60.7%
Top 100 (U.S.)	122,486	67,359	55,127	45.0%
Top 50 (U.S.)	64,381	42,116	22,265	34.6%
Top 20 (U.S.)	26,480	17,834	8,646	32.7%

4.2 NO_x plume detection probabilities have regional disparities

With power plants near interfering emitters removed, we proceeded with a clean set of snapshots in which emissions can be attributed to the target power plant. We apply the pipeline to 500 U.S. plants (January 2019–December 2024) and 6,000 global plants (May–December 2018), excluding power plants within interference zones. Detection frequency varies dramatically across sites.

In the United States (171 after filtering with full 6-year availability), detections range from 10.67 to 243.17 per plant on average each year, with a median of 46.5. This range is highly uneven: plants in the 90th percentile (152.33) record about 8.38× more detections per year than those in the 10th percentile (18.17). When adjusted for the number of satellite overpasses, detection probabilities span from 6.02 to 81.59 per 100 observations (median 20.39 per 100; P10=9.85, P90=71.30).

Globally (1,065 plants after filtering), detections over the eight-month period in 2018 range from 0 to 248, with a median of 37.0. Plants in the 90th percentile (140.7) record about 11.73× more detections than those in the 10th percentile (12.0). When adjusted for the number of satellite overpasses, detection probabilities span from 0 to 100 per 100 observations (median 27.48 per 100; P10=10.25, P90=80.73).



370 These disparities have a clear geographical distribution: detection probabilities are higher in arid and semi-arid regions (e.g., the U.S. interior West and Great Plains, Iberia, North Africa, Greece/Turkey, central Brazil/Argentina, eastern Australia) and lower in humid regions. Notably, the detectability of plants with the same level of emissions varies across different geographical regions, highlighting that emissions alone are insufficient to predict detectability. Regional meteorology and surface properties strongly affect plume detectability as well (Fig. 6a,c), as we will see in the next section.

375 Despite this aggressive filtering, satellite detection rates remain remarkably high (Figure 6). All 171 U.S. plants exhibit detection rates exceeding 10 snapshots per year, with 87.7% detected more than 20 times annually and nearly half (45.0%) detected more than 50 times per year. The global dataset shows similarly strong performance, with 93.0% of plants detected more than 10 times and three-quarters (74.5%) detected more than 20 times during the eight-month observation period. Notably, 23.4% of U.S. plants and 19.4% of global plants exceed 100 positive detections in their respective observation periods, demonstrating
380 TROPOMI's capability for frequent monitoring even after conservative spatial filtering (Figure 7). Detection frequency generally increases with emission magnitude, though substantial variability exists within emission bins, likely reflecting differences in meteorological and environmental conditions across facilities.

4.3 Meteorology, environment, and emissions enable accurate prediction of NO_x plume detectability

Our model predicts plume detectability from meteorological and emissions features with high accuracy (Table 3). Across
385 U.S. and global datasets, the multi-layer perceptron reaches 0.74–0.76 accuracy, 0.65–0.67 F1 score and achieves AUCs above 0.80. These metrics show that sensor, power plant, meteorological, and environmental variables can predict plume detectability, though imperfectly. The residual errors suggest that key drivers are either unmeasured or captured too coarsely, limiting the separability between detectable and undetectable plumes.

Model performance improves substantially when focusing on high-magnitude emitters. As emission strength increases,
390 Precision, Recall, and the F1 score all rise, with Precision showing the most significant improvement. This improvement indicates that the model makes fewer false positive errors when identifying these stronger, less ambiguous plume signals.

(We note that the slight decrease in accuracy in the U.S. when focusing on high emitters is caused by a significant shift in class balance. The full dataset contains a large number of “undetectable” cases (true negatives), and the model correctly identifies them, boosting the overall accuracy score. When the dataset is filtered to only the top emitters, these easier-to-classify negative
395 cases are removed.)

Though the global dataset uses coarser temporal emission data, the model trained on it performs similarly as the model trained on the U.S. dataset. This may initially appear surprising; however, we speculate that the comparable performance might be because the broader variability in environmental variables and emissions in the global dataset sharpens the boundary between detectable and undetectable cases, which in turn helps the model trained on the global dataset to perform better.

400 To understand how model performance varies across the world, we performed a spatial analysis of F1 scores for individual power plants (Figure 8). The geographic distribution reveals distinct regional patterns in model accuracy. For the U.S.-trained model (panel a), performance is strongest in the Midwest and parts of the Western U.S., where the model achieves F1 scores consistently above 0.8 for major emitters. The globally trained model (panel b) demonstrates robust performance across indus-

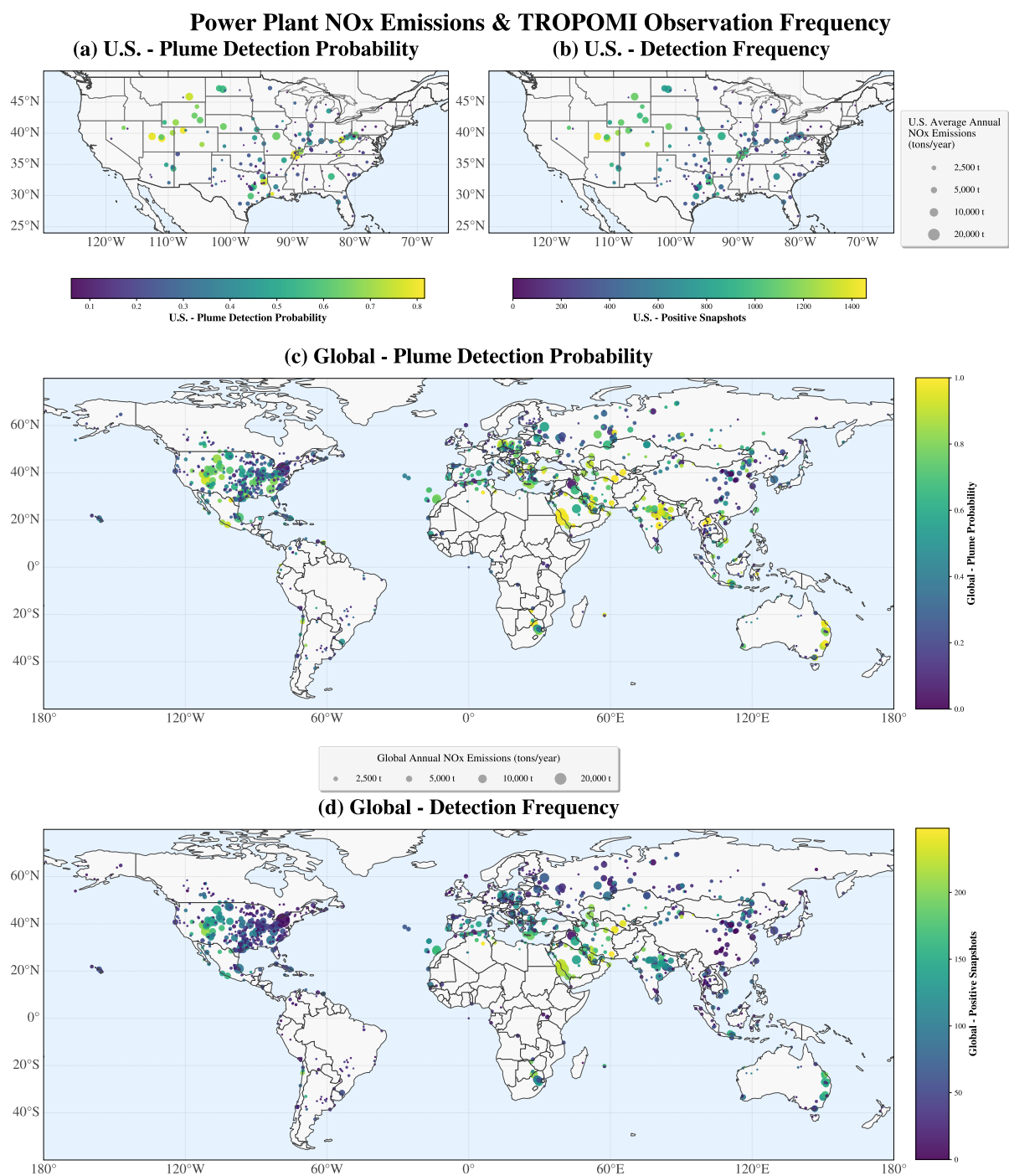


Figure 6. Maps of NO_x emissions and TROPOMI observations: marker size $\propto$ annual NO_x (metric t/year) at each plant, with U.S. panels colored by (a) plume detection probability and (b) positive observation count, and global panels colored by (c) plume probability and (d) positive observation count.



TROPOMI NO_x Plume Detection Frequency vs Annual Emissions - Grid Analysis

(a) United States Power Plants
2019-2024

(b) Global Power Plants
2018

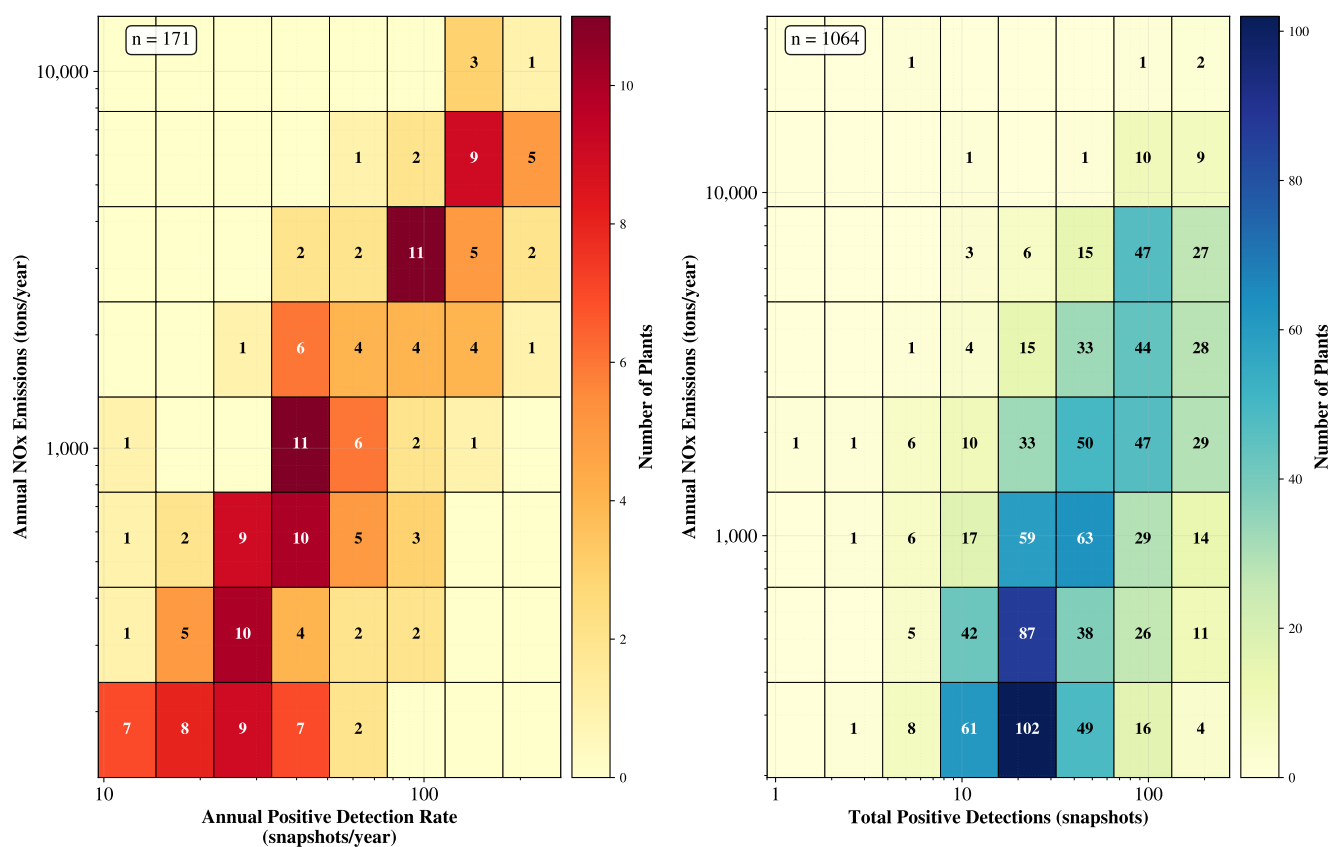


Figure 7. TROPOMI NO_x detection frequency versus annual emissions after interference-zone filtering. (a) U.S. plants ($n = 171$, retaining 49.1% of emissions from the original 500 plants). (b) Global plants ($n = 1065$, retaining 21.1% of emissions from the original 6000 plants). The total number of plants plotted is 1064, as one plant yielded no positive detections. Numbers denote plant counts per grid cell (8×8 log-uniform bins).

**Table 3.** Multi-Layer Perceptron Performance — U.S. and Global

Region	Emissions	Dataset	Accuracy	Precision	Recall	F1 Score	AUC	Kappa
U.S.	Annual NO _x	All	0.750 ± 0.004	0.619 ± 0.009	0.669 ± 0.010	0.643 ± 0.002	0.799 ± 0.003	0.451 ± 0.006
		Top-100	0.707 ± 0.003	0.757 ± 0.012	0.713 ± 0.020	0.734 ± 0.005	0.777 ± 0.002	0.408 ± 0.007
		Top-50	0.702 ± 0.007	0.801 ± 0.010	0.703 ± 0.020	0.749 ± 0.009	0.771 ± 0.006	0.387 ± 0.011
		Top-20	0.710 ± 0.003	0.856 ± 0.005	0.718 ± 0.004	0.781 ± 0.003	0.771 ± 0.009	0.360 ± 0.011
	Hourly NO _x	All	0.763 ± 0.004	0.645 ± 0.012	0.660 ± 0.018	0.653 ± 0.003	0.809 ± 0.002	0.473 ± 0.004
		Top-100	0.733 ± 0.004	0.781 ± 0.008	0.736 ± 0.008	0.758 ± 0.004	0.804 ± 0.003	0.460 ± 0.009
		Top-50	0.734 ± 0.003	0.816 ± 0.006	0.747 ± 0.015	0.780 ± 0.006	0.805 ± 0.003	0.445 ± 0.006
		Top-20	0.739 ± 0.009	0.870 ± 0.006	0.749 ± 0.019	0.805 ± 0.010	0.805 ± 0.004	0.415 ± 0.012
Global	Annual NO _x	All	0.745 ± 0.005	0.679 ± 0.010	0.652 ± 0.007	0.665 ± 0.003	0.800 ± 0.003	0.459 ± 0.008
		Top-100	0.755 ± 0.006	0.890 ± 0.007	0.764 ± 0.013	0.822 ± 0.005	0.828 ± 0.009	0.436 ± 0.018
		Top-50	0.793 ± 0.020	0.893 ± 0.016	0.814 ± 0.018	0.852 ± 0.016	0.864 ± 0.020	0.509 ± 0.040
		Top-20	0.850 ± 0.031	0.963 ± 0.013	0.859 ± 0.040	0.907 ± 0.020	0.902 ± 0.036	0.522 ± 0.090

trialized regions including North America, Europe, and the Arabian Peninsula, with F1 scores frequently exceeding 0.8–1.0 for large point sources. Performance degrades in parts of Asia, South America, and Africa, where F1 scores show greater variability and often drop below 0.5, likely because training data are sparser and emissions are lower in these regions. Across all regions, a clear positive relationship exists between emission magnitude (point size) and model performance: larger NO_x emitters are detected with higher accuracy, likely because stronger plume signals exceed detection thresholds more reliably (verified in Figure C2).

One notable outlier appears in the northeastern U.S. The global model shows poor performance for what appears to be the largest emitter in the U.S. after interference zone filtering in the COCO₂ dataset, yet this facility is absent from the U.S. analysis. This facility, Domtar Paper Company, LLC, reports 2018 NO_x emissions of only 35.57 tons in the EPA dataset but 29,652 tons in COCO₂—an 800-fold discrepancy. If the EPA data is correct, the facility’s small actual size explains the poor detectability. In other words, the COCO₂ dataset is likely incorrect about Domtar Paper Company’s emissions.

4.4 Which features are most predictive of plume detectability?

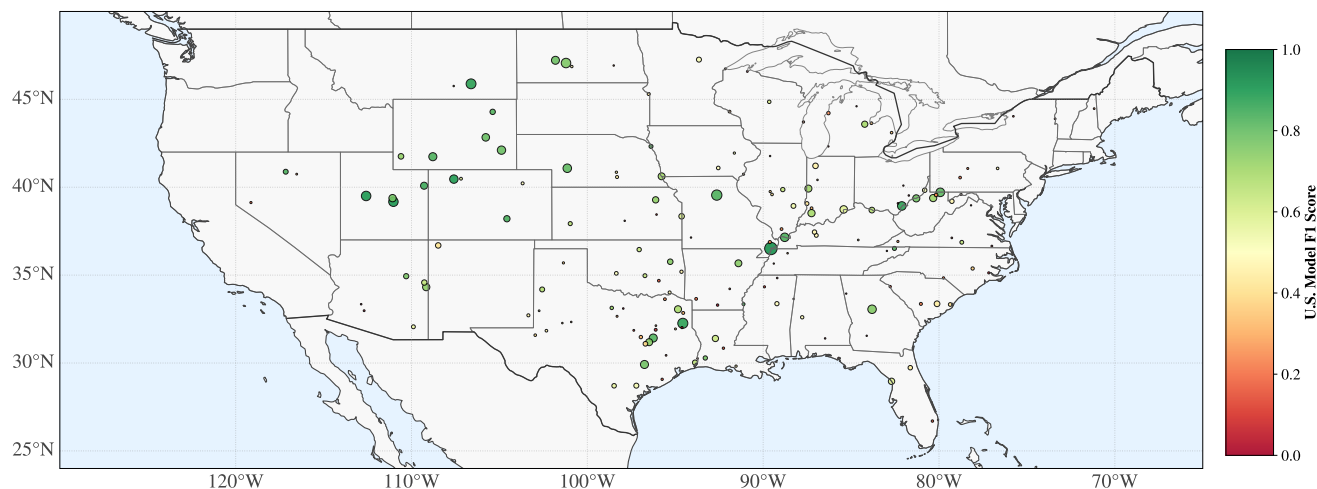
To identify the key predictors of NO₂ plume detectability, we performed a permutation importance analysis on our best-performing models (Table 3). The top 5 environmental and power plant-specific features include NO_x emission rate, surface albedo, sensor zenith angle, primary fuel type, and wind speed (Figure 9, left). Our SHAP score analysis yielded similar results, which are shown in Appendix C1. A complete list of permutation-importance values appears in Appendix E, Table E1.

To examine how these features shape detectability, we employ a quantile-based binning strategy, using data between the 1st and 99th percentiles. This approach ensures a balanced distribution of samples across bins while mitigating the influence of extreme outliers. We grouped the data into these quantile bins for each feature ($\approx 7,800$ samples/bin globally; $\approx 9,700$ in the U.S.) and computed the probability of detection within each bin. While we acknowledge that these variables can be correlated, we find the trends persist when we parse the data to isolate each feature’s effect (e.g., by holding other variables roughly constant). The resulting curves are plotted against the corresponding feature values (Figure 9, right).



Per-Plant Model Performance: F1 Scores from MLP Models

(a) U.S. - Model F1 Score



(b) Global - Model F1 Score

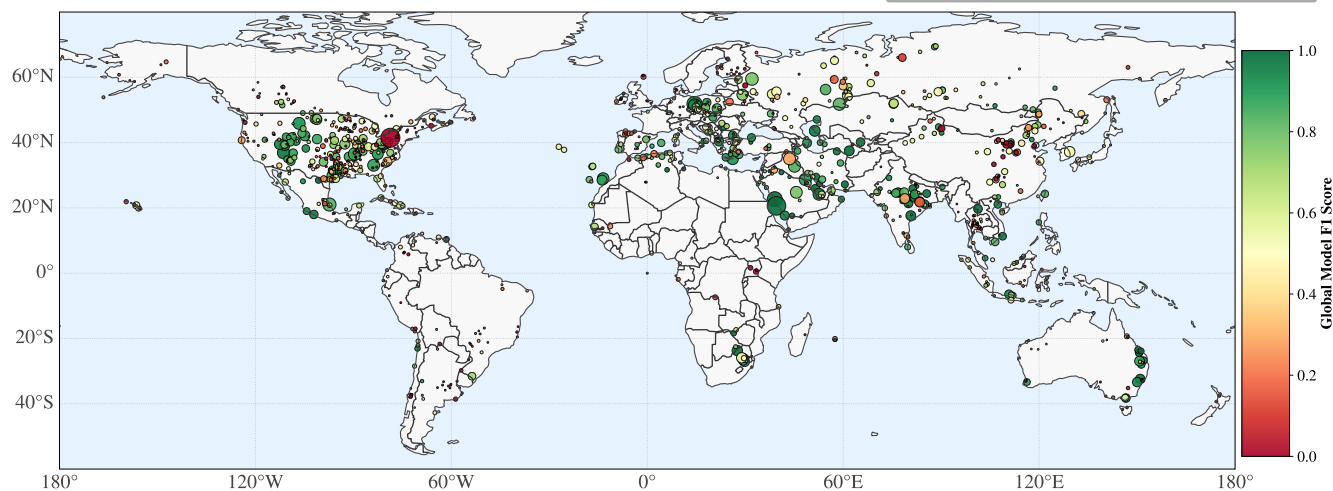


Figure 8. Geographic distribution of multi-layer perceptron (MLP) model performance in detecting NO_x plumes from individual power plants, evaluated using the F1 score. Panel (a) shows results for the U.S. model, while panel (b) presents the global model. Each point represents a power plant, with color indicating predictive performance on a scale from 0.0 (poor) to 1.0 (perfect), and marker size proportional to annual NO_x emissions. We can see a general trend that larger emitters tend to have higher F1 scores. The U.S. model demonstrates strong performance over mid-west areas. The global model demonstrates strong performance across North America, Europe, and the Arabian Peninsula.



Feature Importance & Probability of Detection Analysis

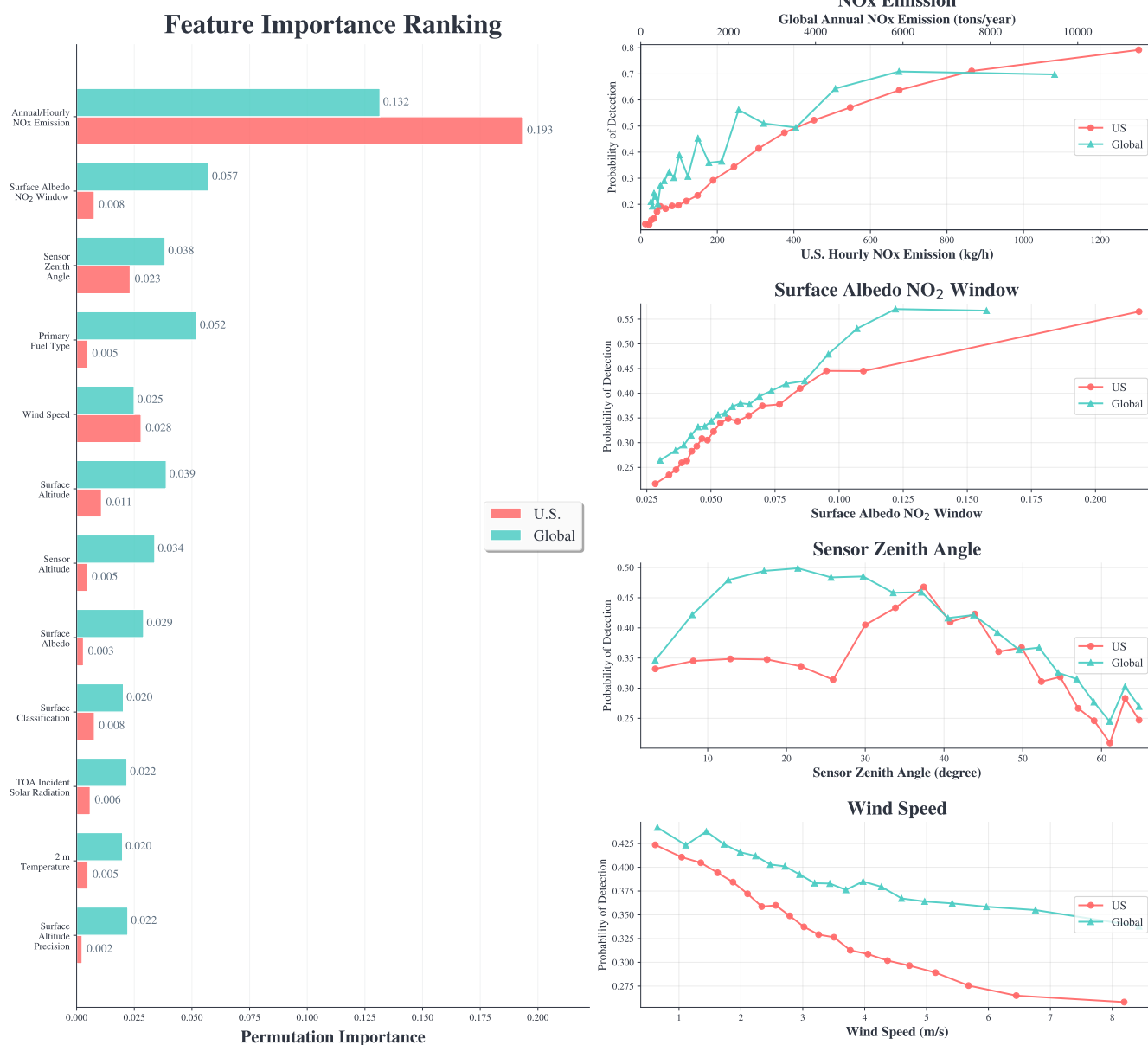


Figure 9. Feature importance and probability of detection analysis for TROPOMI-based NO₂ models at U.S. (red) and global (teal) scales. Left: ranked feature importance. Right: detectability as a function of the top three features and wind speed. Binning uses a quantile-based strategy (1st–99th percentile) to ensure balanced feature distribution and avoid extreme values.



Detectability increases with both emission rate and surface albedo. As expected, higher emissions increase NO_2 concentration and therefore the detectable signal, while brighter surfaces increase scene radiance, improving signal-to-noise ratio (SNR). Our analysis finds that for U.S. power plants, an average hourly emission rate of ≈ 400 kg/h corresponds to a 50% detection probability. The result for global power plants is similar, although the global $P(\text{detection})$ versus NO_x emission curve is slightly above the U.S. curve. To compare U.S. and global performances in Figure 9, we align their emission axes to share the same numerical scale, despite the different units (kg/h for the U.S., tons/year for global). We align the x-axis for U.S. emissions (kg/h) and global emissions (tons/year) by converting the U.S. hourly rates to annual values using $24 \times 365/1000$, allowing direct comparison with the global annual rates. The U.S. curve consistently falling below the global curve suggests that the premise of continuous operation is often invalid, as many U.S. plants run intermittently, leading to lower detection probabilities.

By contrast, the sensor zenith angle (SZA) shows a hump-shaped dependence. From nadir to moderate off-nadir, detectability rises because longer slant paths at moderate SZA increase the tropospheric air-mass factor and the line of sight intersects more of the plume, boosting plume-background contrast (Palmer et al., 2001; Eskes and Boersma, 2003). At large SZA, detectability falls as radiance drops and multiple scattering grows, degrading SNR; retrieval uncertainties also increase under oblique geometry (Van Geffen et al., 2020). In our data, the average mean peak plume detectability occurs near 20° SZA globally and 35° in the U.S.

Finally, wind speed is negatively correlated with detectability. Stronger winds advect and dilute plumes, lowering column contrast and thus the probability of detection. We note that this is different from the ideal wind speeds for quantifying emissions, which prior literature shows to occur at moderate wind speeds (2–4 m/s). Very low wind speeds make it difficult to quantify the rate of emission due to the high relative uncertainty in the wind measurement itself, while the same plume dilution that hinders detection at very high wind speeds also introduces large errors into the quantification algorithm (Bruno et al., 2024).

Primary Fuel Type is a significantly less important feature in the U.S. model (feature importance: 0.005) compared to the global model (feature importance: 0.052). This difference likely stems from the relative homogeneity of the U.S. power sector in both its fuel sources and emission controls. First, the U.S. cohort of top-emitting power plants is dominated by natural gas ($\approx 61\%$) and coal ($\approx 32\%$), which together account for $\approx 93\%$ of the facilities. This low fuel variability is in stark contrast to the global dataset. While also led by coal ($\approx 50\%$) and natural gas ($\approx 30\%$), the global mix is far more diverse, including substantial shares of facilities running on oil ($\approx 12\%$), biomass ($\approx 6\%$), and waste ($\approx 2\%$). Second, this effect is compounded by emission controls. In the U.S., power plants are subject to relatively uniform and stringent environmental regulations, such as the Clean Air Act, often requiring technologies like Selective Catalytic Reduction to control NO_x emissions. Globally, however, the presence and effectiveness of these technologies vary drastically by country. In that context, fuel type can serve as a stronger proxy for a facility's overall emission profile.

While the feature importance analysis identifies the key predictors and their marginal effects, real-world plume detectability emerges from the combined influence of multiple interacting factors. To illustrate how emission rates, albedo, and wind speed operate in concert at individual facilities, we examine five representative power plants that span the range of conditions observed in our dataset (Figure 10). Richmond County Plant, with low NO_x emissions (~ 72 kg/h) and low albedo (0.04), has the lowest



detection probability ($P(\text{detect}) = 0.13$). Ottumwa shows moderate NO_x (~ 185 kg/h) but experiences high wind speeds (4.24 m/s) and variable albedo (0.07 ± 0.08), reflected in its long-tail distribution from winter snow cover, yielding $P(\text{detect}) = 0.24$.

Comparing Richmond County and Independence directly demonstrates the dominant effect of emission rate: Independence's high NO_x (~ 662 kg/h) achieves $P(\text{detect}) = 0.56$ despite similarly low albedo (0.05). Among the three high-emitting plants
465 (Independence, Dave Johnston, Bonanza), surface characteristics become the limiting factor. Independence's low albedo con-
strains detectability to 0.56, while Dave Johnston (high wind, 5.01 m/s) reaches 0.63, and Bonanza (low wind, 2.74 m/s; high
albedo, 0.13) achieves the highest detection probability at 0.73. These examples demonstrate that high surface albedo and low
wind enhance NO_2 retrieval sensitivity even when emission rates are comparable.

5 Discussion

470 This study provides the first global, data-driven quantification of when and where power-plant NO_2 plumes are visible from
space. Previous satellite analyses of power plants have typically examined a handful of large facilities under selected conditions
to demonstrate detectability or to infer emissions (Goldberg et al., 2019; Beirle et al., 2019; Liu et al., 2020; Cusworth et al.,
2021). In contrast, we analyze more than 6,000 facilities worldwide using a consistent, automated plume-labeling and machine-
learning framework that links satellite detectability to meteorological, environmental, and sensor-geometry features. Our goal
475 in establishing empirical relationships governing plume visibility is to elucidate the current limits of satellite NO_2 sensing and
inform future satellite-based NO_x emission quantification.

5.1 Empirical characterization of plume detectability

Our results empirically confirm several theoretical expectations from satellite trace-gas retrieval physics while quantifying their
combined effects on plume visibility at global scale. Viewing geometry exerts a strong nonlinear influence: moderate off-nadir
480 angles enhance detectability by increasing the optical path length through the lower troposphere, consistent with established
air-mass-factor and averaging-kernel theory (Palmer et al., 2001; Eskes and Boersma, 2003). At larger angles, detectability
declines as scene radiance decreases and multiple scattering increases (Van Geffen et al., 2020). Surface reflectance also plays
a major role, with higher albedo increasing scene radiance and retrieval SNR, improving the contrast between a plume and its
background, consistent with TROPOMI characterization studies (Van Geffen et al., 2020).

485 Our analysis extends these theoretical principles by quantifying their global impact empirically. Across thousands of facili-
ties, detectability peaks at intermediate sensor zenith angles of 25 to 30° and increases monotonically with both emission rate
and surface brightness. For the U.S. subset, an average hourly NO_x emission rate of roughly 400 kg h^{-1} corresponds to a 50%
probability of detection. These results provide the first global observational confirmation of how viewing geometry, surface
brightness, meteorology, and emission strength jointly control satellite plume visibility.

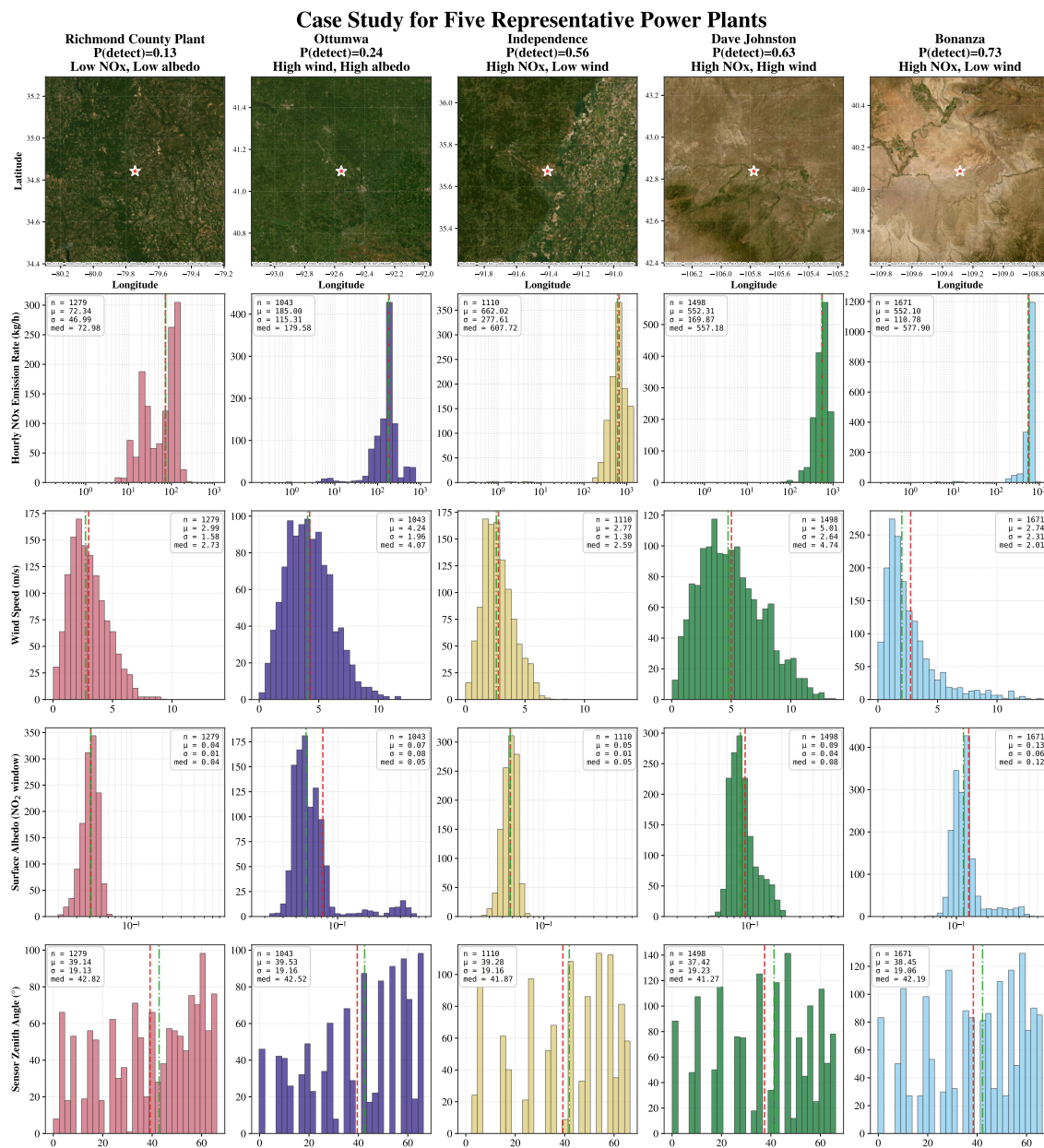


Figure 10. Variable distributions for five representative power plants demonstrating detectability drivers. Plants were selected from 171 consistently monitored power plants in the U.S. to showcase diverse NO_x emissions, meteorological conditions, and viewing geometries. Top: Satellite imagery with detection probabilities and dominant characteristics. Rows 2-5: Histograms of hourly NO_x emission rates (kg/h), wind speed (m/s), surface albedo (NO₂ window), and sensor zenith angles, with mean (red dashed) and median (green dash-dot) lines. Consistent scales enable cross-plant comparison. Satellite Image Source: Esri, i-cubed, USDA, USGS, AEX, GeoEye, Getmapping, Aerogrid, IGN, IGP, UPR-EGP, and the GIS User Community | Powered by Esri.



490 5.2 Implications for emission retrieval

The same meteorological, environmental, and geometric variables that determine whether a plume is detectable also shape how it appears in satellite imagery. Wind speed controls plume width and dispersion, solar and sensor geometry set the radiative contrast and effective path length, and surface albedo defines the background brightness against which NO₂ enhancements are retrieved. If these factors affect visibility, then they also influence the quantitative mapping between observed plume morphology/enhancement and underlying emissions. Consequently, emission-retrieval frameworks — whether physics-based inversions or machine-learning models — are likely to benefit by explicitly incorporating these variables. In inverse modeling, detectability probabilities or their underlying covariates could serve as weights or priors, emphasizing scenes with favorable observing conditions and reducing biases introduced by low-sensitivity observations. In ML-based retrievals, including features such as wind speed, solar/sensor angles, and surface albedo alongside the imagery would allow models to learn how identical plume appearances can correspond to different true emissions under different conditions. This contextual information should improve generalization across power plants and climates, reduce bias in data-poor regions, and produce more physically consistent emission estimates.

6 Conclusions

In this work, we systematically mapped power plant NO₂ plume detectability at U.S. and global scales, and then demonstrated that detectability can be predicted by a suite of environmental and sensor variables. In so doing, we are the first to demonstrate the large variance in plume detectability due largely to the interaction of NO_x emission level, surface albedo, wind speed, and sensor zenith angle. Our results empirically validate long-standing theoretical expectations from satellite retrieval physics while offering the first statistical measure of how these factors combine to govern plume appearance in real observations.

Despite its scope, this analysis has several limitations. A large fraction of power plants, especially in densely industrialized or urban regions, were excluded due to proximity to other emission sources. We opted for conservative filtering to reduce confounding from overlapping sources in our second-stage analysis of plume visibility. However, removing a large percentage of power plants from analysis does potentially limit the representativeness of our analysis in regions where interference is common. Future work could incorporate wind direction or source separation techniques to model overlapping plumes more explicitly. In addition, some variables that plausibly influence detectability were unavailable from TROPOMI or ERA5, including vertical wind shear, aerosol optical depth, and more detailed measures of turbulence. Finally, while this study predicts whether plumes are detectable, the next critical step is to use the same environmental and observational predictors that govern plume visibility to quantify emissions themselves, accounting for how these factors also shape the relationship between plume appearance in satellite imagery and the underlying emission rate.

Code and data availability. Data used in the paper are available on Zenodo (<https://doi.org/10.5281/zenodo.17706582>). The code will be open-sourced when accepted.



Appendix A: Annual Interference Filtering Statistics for U.S. Top-500 Emitters

Table A1. Annual TROPOMI snapshot filtering for U.S. top-500 NO_x emitters (2019–2024). The filtering removes snapshots where target plants are within interference zones from nearby urban areas or other major power plants.

Year	Unfiltered	Retained	Removed	Removal Rate
2019	84,658	33,453	51,205	60.5%
2020	82,795	32,517	50,278	60.7%
2021	82,063	32,023	50,040	61.0%
2022	86,676	33,723	52,953	61.1%
2023	82,745	32,631	50,114	60.6%
2024	85,228	33,991	51,237	60.1%
Total	504,165	198,338	305,827	60.7%

The filtering statistics show remarkable temporal consistency, with removal rates stable at 60–61% across all years. This consistency reflects the relatively static nature of interference zones: urban boundaries and power plant configurations change slowly compared to the annual observation cycle. The retained 198,338 snapshots (39.3% of the original dataset) represent observations where plumes could be attributed to individual plants with high confidence, enabling causal analysis of how meteorological, environmental conditions, sensor geometry, and emission rates affect plume detectability.

Appendix B: Model Training Details

To ensure fully reproducible and balanced training, we first partitioned the dataset into a training set (60%), a validation set (20%), and a test set (20%), using a fixed random state (`random_state=42`) for deterministic shuffling. We then applied random oversampling to the minority class in the training set to address class imbalance. Subsequently, all features were standardized to zero mean and unit variance by fitting a `StandardScaler` from `sklearn.preprocessing` on the resampled training data and applying the same transformation to the validation and test sets.

For the model itself, weights were initialized uniformly from the range $[-1/\sqrt{\text{in_features}}, 1/\sqrt{\text{in_features}}]$. We trained the model for up to 100 epochs using the Adam optimizer with a batch size of 32 and a binary cross-entropy with logits loss function. We performed a search over seven learning rates— 1×10^{-5} , 2×10^{-5} , 5×10^{-5} , 1×10^{-4} , 2×10^{-4} , 5×10^{-4} , 1×10^{-3} , selecting the configuration with the highest AUC (Area Under the Curve) on the validation set for final evaluation on the test set. An early-stopping rule with a patience of five halted training if the validation loss did not improve for five consecutive epochs, after which the best-performing model was restored. To account for training stochasticity, we ran each experiment five times and reported the mean and variance of the results.



540 **Appendix C: Complementary Results for Item Split Experiments**

C1 Plant-level Feature Importance via SHAP Score

For the U.S. dataset, across the six-year record, 171 power plants never fall within interference regions. While training for U.S. models, we maximize the number of usable snapshots to ensure better training, but for analysis, we restrict attention to these 171 plants.

545 SHAP Score Analysis confirms the permutation results—NO_x emissions dominate globally—while adding plant-level nuance. We map the top feature per plant (Fig. C1a,c) and the regional modal top feature on a grid (Fig. C1b,d).



Plant-level Feature Importance via SHAP Score

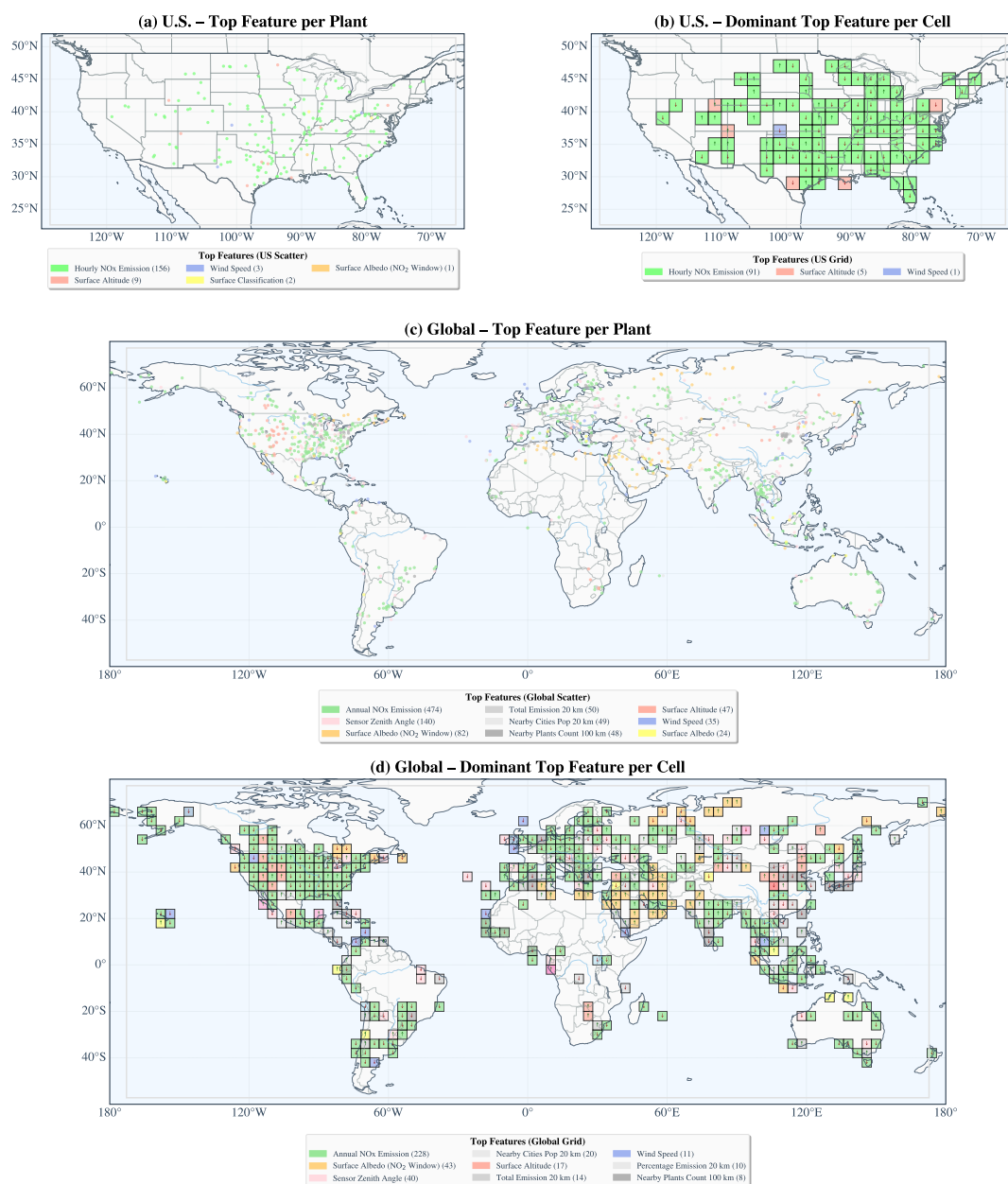


Figure C1. Dominant predictive features of plume detectability across the U.S. and globally (predictors include nearby-city and nearby-plant statistics). **(a,c)** Plant-level scatter: each point is a power plant colored by its most influential feature. **(b,d)** Grid-mode maps: within each cell (2° U.S.; 4° global), the color marks the modal top feature; arrows indicate positive (↑) or negative (↓) effects. Colors are consistent across panels.



C2 Correlation between Per-Plant Model Performance and Key Features

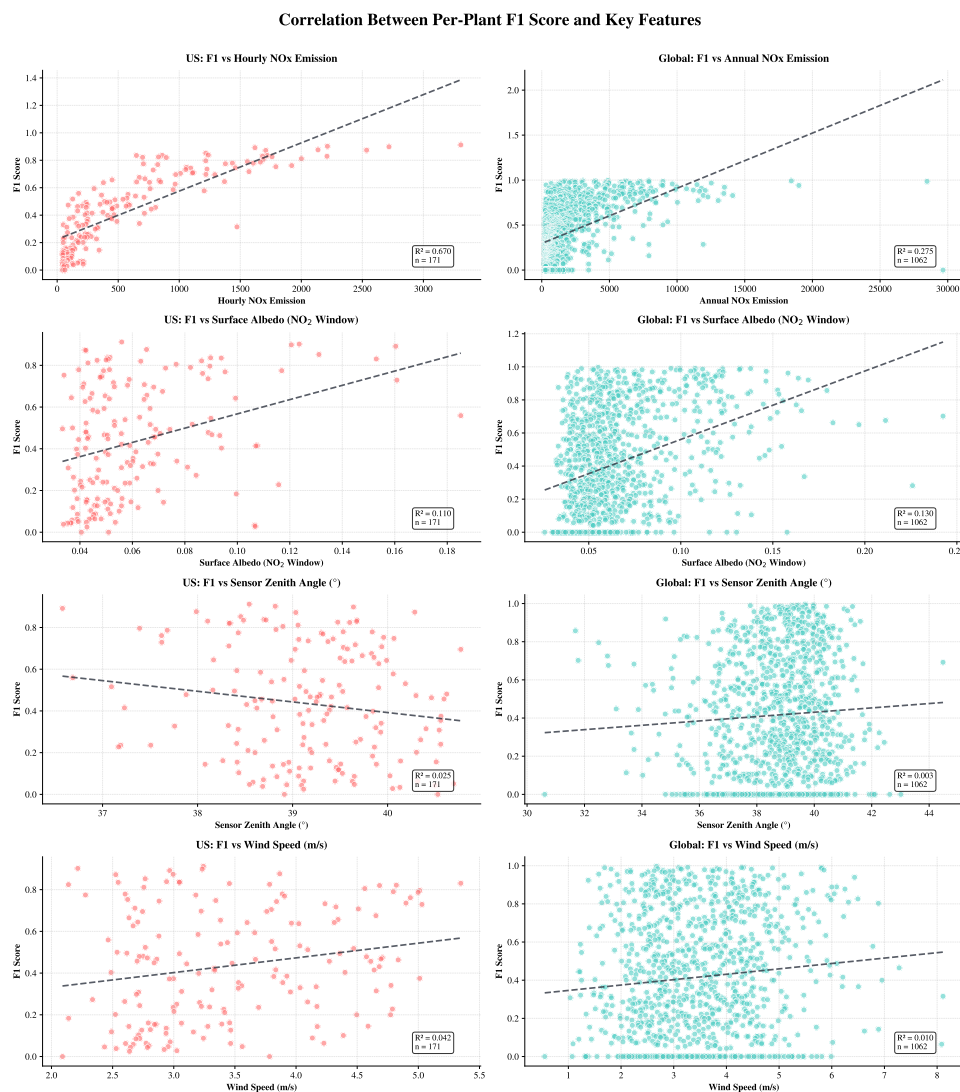


Figure C2. Correlation between F1 Score and Key Features for U.S. and Global Power Plants. Scatter plots show the relationship between the plume detection F1 score and four key features, separated for the U.S. dataset (left column, $n=171$) and a Global dataset (right column, $n=1062$). We filtered out power plants with less than 3 snapshots, and that's why the available power plants for global analysis drop from 1065 to 1062. The features (aggregated as plant-level means) are, from top to bottom: NO_x emissions, surface albedo, sensor zenith angle, and wind speed. Each point represents an individual power plant. The dashed line in each panel indicates the linear regression fit, with the corresponding coefficient of determination (R^2) and sample size (n) displayed. The strongest positive correlation is observed between the F1 score and hourly NO_x emissions for both datasets, especially in the U.S. ($R^2 = 0.670$). All other features exhibit significantly weaker correlations.



C3 Probability of Detection Analysis with Two Features

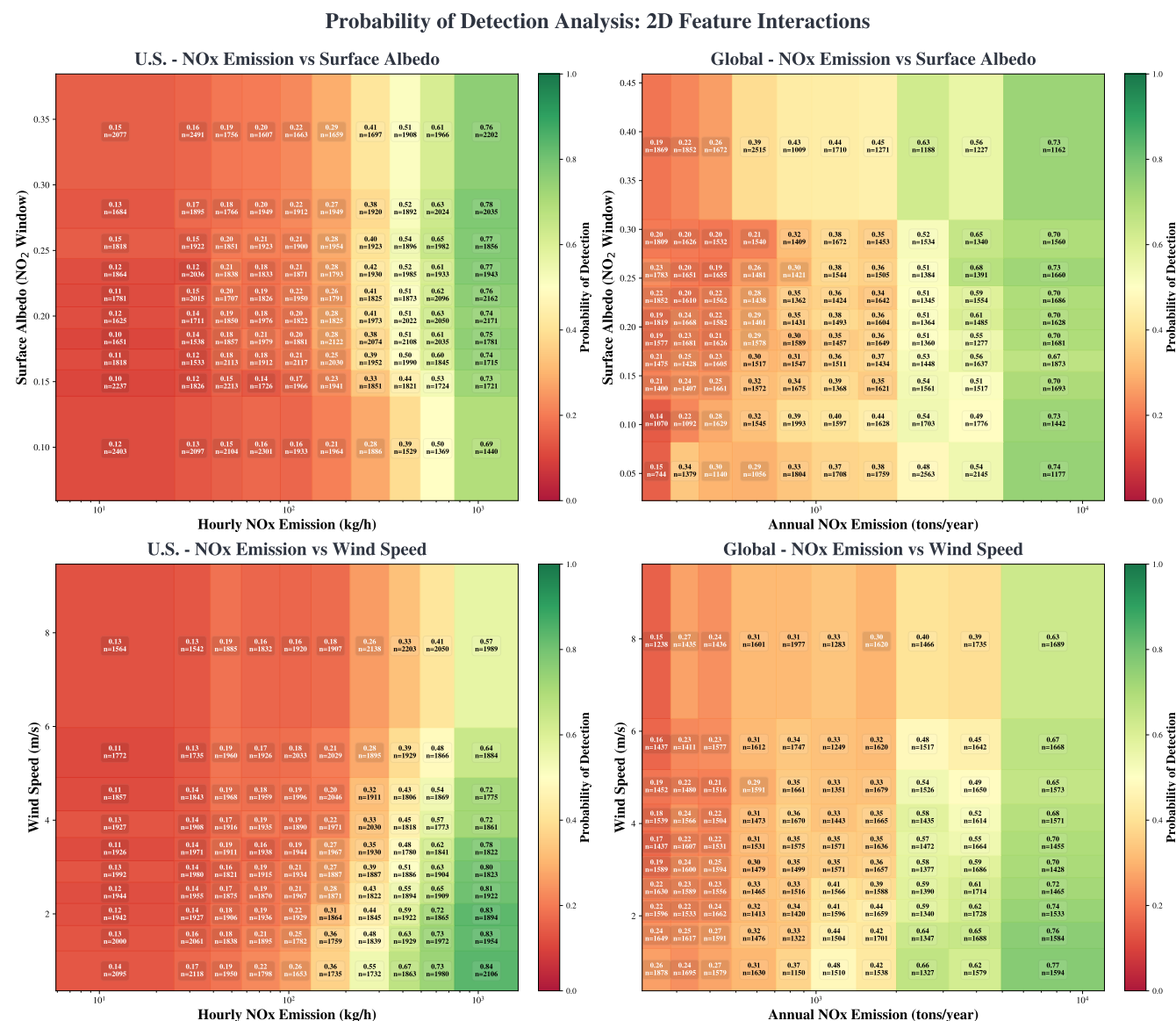


Figure C3. Two-dimensional probability of detection (POD) analysis for TROPOMI NO₂ plume detection from power plants. Heatmaps show POD as a function of NO_x emission rate and key environmental variables for U.S. facilities (left column; hourly emission in kg/h) and global facilities (right column; annual emission in tons/year). Top row: NO_x emission vs. surface albedo in the NO₂ spectral window. Bottom row: NO_x emission vs. 10-m wind speed. Color intensity indicates POD (red = low, green = high); cell annotations display POD values and sample counts (n). Binning uses a quantile-based strategy (1st–99th percentile) to ensure balanced feature distribution and avoid extreme values.



Appendix D: Power Plant Split Experiments

In the item-based split, individual observations are randomly assigned to the training, validation, and test sets, yielding an in-distribution estimate because observations from the same plant can appear in multiple sets (see Section 4.3). To test the generalization across power plants, we perform the experiment of power plant split.

After applying the interference-zone filter, 183 power plants have at least one clean year of data. We retained 183 from Top 500 U.S. plants (52 from Top 100, 32 from Top 50) and 1,035 from Top 6000 global plants (35 from Top 100, 18 from Top 50) with ≥ 1 year of usable data. The power plant split allocated entire power plants to each set, ensuring the test set contained only previously unseen facilities. We first excluded those power plants within the interference zone, then split the remaining plants 60/20/20 across training, validation, and test sets.

The power plant split results demonstrate strong generalization performance to unseen power plants, validating that the learned features capture transferable emission patterns rather than facility-specific signatures. Comparing Tables 3 and D1 reveals the model's ability to generalize across different facilities.

Robust Generalization to New Facilities. The features demonstrate robust generalization to unseen power plants with modest performance declines for comprehensive datasets. In the U.S., accuracy drops only 2–7% for the “All” dataset (0.750→0.697 for annual NO_x , 0.763→0.742 for hourly NO_x), while global accuracy declines 6.7% (0.745→0.695). AUC scores remain strong across all scenarios, indicating the model maintains good discriminative ability when tested on entirely new facilities.

The model's performance declines more significantly when tested on subsets of high-emitting plants. This drop is likely because of the small overall training dataset, which provides insufficient data for the model to generalize effectively to facilities with unique operational signatures.

The results confirm that learned features capture generalizable emission signatures rather than memorizing power plant-specific characteristics. While performance naturally declines when testing on entirely new plants—particularly for high-emitting subsets—the maintained discriminative ability across diverse facilities validates the robustness of our approach for real-world deployment scenarios.

Table D1. Multi-Layer Perceptron Performance (Power Plant Split) — U.S. and Global

Region	Emissions	Dataset	Accuracy	Precision	Recall	F1 Score	AUC	Kappa
U.S.	Annual NO_x	All	0.697 $\pm$ 0.003	0.569 $\pm$ 0.005	0.710 $\pm$ 0.010	0.632 $\pm$ 0.002	0.772 $\pm$ 0.002	0.380 $\pm$ 0.003
		Top-100	0.626 $\pm$ 0.004	0.721 $\pm$ 0.006	0.522 $\pm$ 0.019	0.605 $\pm$ 0.011	0.657 $\pm$ 0.003	0.267 $\pm$ 0.006
		Top-50	0.675 $\pm$ 0.007	0.745 $\pm$ 0.009	0.708 $\pm$ 0.029	0.726 $\pm$ 0.012	0.721 $\pm$ 0.005	0.329 $\pm$ 0.008
	Hourly NO_x	All	0.742 $\pm$ 0.003	0.631 $\pm$ 0.006	0.712 $\pm$ 0.005	0.669 $\pm$ 0.001	0.801 $\pm$ 0.002	0.459 $\pm$ 0.005
		Top-100	0.657 $\pm$ 0.005	0.756 $\pm$ 0.012	0.557 $\pm$ 0.031	0.640 $\pm$ 0.016	0.708 $\pm$ 0.013	0.327 $\pm$ 0.007
		Top-50	0.717 $\pm$ 0.006	0.769 $\pm$ 0.015	0.764 $\pm$ 0.017	0.766 $\pm$ 0.002	0.769 $\pm$ 0.007	0.408 $\pm$ 0.018
Global	Annual NO_x	All	0.695 $\pm$ 0.005	0.596 $\pm$ 0.011	0.538 $\pm$ 0.018	0.565 $\pm$ 0.007	0.730 $\pm$ 0.004	0.331 $\pm$ 0.008
		Top-100	0.588 $\pm$ 0.020	0.851 $\pm$ 0.015	0.583 $\pm$ 0.034	0.691 $\pm$ 0.022	0.623 $\pm$ 0.017	0.129 $\pm$ 0.029
		Top-50	0.616 $\pm$ 0.044	0.918 $\pm$ 0.010	0.623 $\pm$ 0.063	0.740 $\pm$ 0.039	0.646 $\pm$ 0.032	0.086 $\pm$ 0.013



Appendix E: Feature Importance and Correlation

Figure E1 shows the correlation between different variables we analyze in the U.S. Dataset and Global Dataset after filtering out the power plants in interference zones. The strongest correlations in both the U.S. and Global datasets reflect a handful of underlying physical and spatial relationships.

First, atmospheric variables such as surface altitude and surface pressure exhibit an almost perfect negative correlation ($\rho = -0.994$ in the U.S., $\rho = -0.986$ globally). This simply mirrors the barometric law: as you ascend, the weight of the air above you diminishes, driving pressure down.

Likewise, solar zenith angle and total incoming solar radiation (TISR) are very tightly linked ($\rho = -0.974$ in the U.S., $\rho = -0.963$ globally): when the sun is closer to the horizon (large zenith angle), less radiation reaches a horizontal surface.

One additional U.S.-specific pattern links cloud cover to surface brightness: cloud fraction and scene albedo correlate at $\rho = 0.865$, as more cloud cover generally leads to brighter reflectance in visible wavelengths.

In all of the above cases, two highly correlated features form a pair and you only need one feature within each pair to capture the same information in a model. Consequently, any pair with $|\rho| > 0.8$ risks multicollinearity in regression models. Future work can mitigate this risk by dropping one feature from each pair or by applying dimensionality-reduction techniques (e.g., principal component analysis) to extract orthogonal components.

Table E1 records the permutation importance across four model configurations we displayed in Section 4.4. The bar chart visualization of Table E1 is Figure 9.

Appendix F: Case Studies in Automated Plume Detection

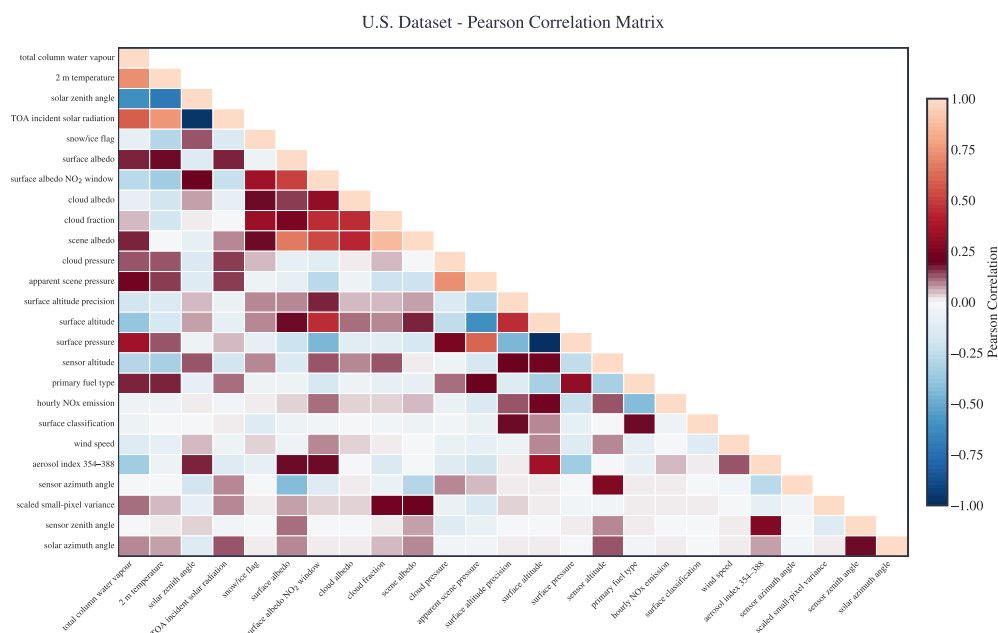
To assess the performance of our Automated Plume Detection Algorithm (Section 3.1), we analyze its application across several detection scenarios. The algorithm's objective is to determine if a power plant's plume is clearly visible and attributable to its source in a given satellite image.

A successful classification results in either a **True Positive**, where a genuine plume is correctly identified and attributed, or a **True Negative**, where the absence of a plume is correctly noted. Conversely, misclassifications fall into two categories:

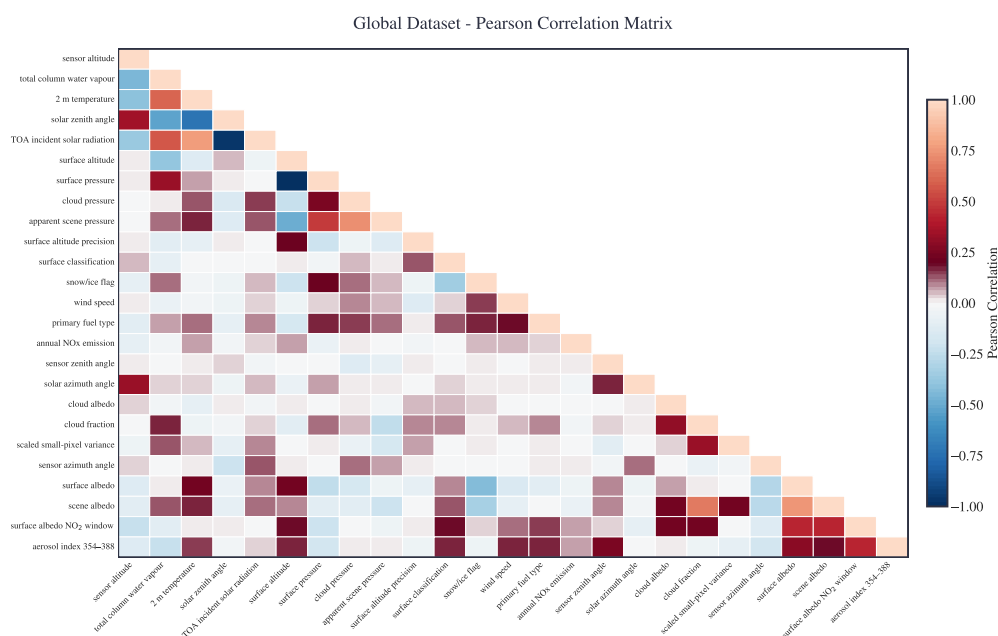
1. **False Positives:** The algorithm incorrectly attributes emissions from other sources to the target power plant.
2. **False Negatives:** The algorithm fails to detect a genuine plume from the target plant.

Minimizing false positives is our primary objective. We address this by applying an *interference mask* to exclude signals from non-target sources. False negatives are inherently more difficult to eliminate because overlapping emissions can be hard to disentangle. Furthermore, the outcome is highly sensitive to the detection threshold and spatial range; even with sensible parameters, some obvious plumes may be missed.

This section moves beyond simple error labels to explore the underlying causes of misclassification, which include:



(a) Feature correlation matrix for U.S. power plant NO_x emissions



(b) Feature correlation matrix for global power plant NO_x emissions

Figure E1. Pearson correlation matrices for TROPOMI-based NO_x emission estimate features. (a) Correlation structure for U.S. power plants. (b) Correlation structure for global power plants. Values range from -1 (strong negative correlation) to +1 (strong positive correlation).



Feature	U.S.	Global
<i>Sensor Features (TROPOMI)</i>		
sensor zenith angle	0.0232	0.0384
sensor azimuth angle	0.0038	0.0071
sensor altitude	0.0046	0.0339
scaled small-pixel variance	0.0034	0.0050
<i>Power Plant Features</i>		
annual NO _x emissions	—	0.1316
hourly NO _x emissions	0.1934	—
primary fuel type	0.0048	0.0521
<i>Meteorology Features</i>		
<i>TROPOMI</i>		
10 m wind speed	0.0280	0.0250
cloud albedo	0.00035	0.00024
cloud pressure	0.0032	0.0039
cloud fraction	0.0029	0.0097
surface pressure	0.0030	0.0209
solar zenith angle	0.0025	0.0169
solar azimuth angle	0.0014	0.0088
apparent scene pressure	0.0038	0.0130
aerosol index 354–388	0.0028	0.0088
<i>ERA5</i>		
2 m temperature	0.0049	0.0199
total column water vapour	0.0034	0.0095
TOA incident solar radiation	0.0059	0.0218
<i>Environment Features</i>		
<i>TROPOMI</i>		
surface classification	0.0077	0.0203
snow/ice flag	0.0021	0.0087
scene albedo	0.0027	0.0134
surface albedo	0.0030	0.0290
surface albedo (NO ₂ window)	0.0076	0.0574
surface altitude	0.0108	0.0389
surface altitude precision	0.0024	0.0222

Table E1. Permutation importance for U.S. and Global models. “—” marks a feature that is absent. Values rounded to 4 decimals.



- **Interference:** Signals from nearby industrial facilities or urban centers are confused with the target plume. Conversely, the interference mask itself can sometimes obscure a genuine plume.
- **Incorrect Threshold:** The detection threshold is either too high, causing faint but real plumes to be missed, or too low, causing background noise to be flagged as a plume.
- **Regional Gradients:** Natural variations in background NO₂ levels, such as the sharp contrast between land and sea, are misinterpreted as an emission plume.

The figures below provide a standardized, six-panel visualization to analyze each detection case. Together, these panels illustrate the NO₂ measurements, plume detection results, background estimation, and geographic context.

- **Top Row (Detection):** The **top-left panel** shows the raw NO₂ concentration field, with the target facility (red cross), nearby industrial sources (blue triangles), and urban centers (yellow dots) marked. The **top-middle panel** overlays the detected plume contour and a wind-direction wedge. The **top-right panel** provides a satellite image for real-world context.
- **Bottom Row (Analysis):** The **bottom-left panel** displays the calculated background NO₂ field. The **bottom-middle panel** maps concentration anomalies relative to this background, highlighting enhancements (warm colors) and showing any masked interference zones. Finally, the **bottom-right panel** uses an OpenStreetMap basemap to illustrate the local road network.

A consistent legend is used across all panels to clarify how plumes are identified, isolated from background signals, and interpreted within their geographical setting.

F1 Successful Detections

F1.1 True Positives: Correctly Identified Plumes

A true positive occurs when the algorithm successfully identifies a distinct plume originating from the target facility. These cases confirm the method's ability to separate a clear emission signal from background concentrations. Figure F1 shows an ideal detection, while Figure F2 demonstrates a successful detection of a weaker, more diffuse plume.



Facility: CoCO2_00224 (AUS) | 2018-09-23T03:43:19.575000Z
Annual NOx: 4885.6 t/yr | Wind: 153.6° @ 2.89 m/s | Plume Detected (Area 173.6 km², Thresh 1.90e-05)

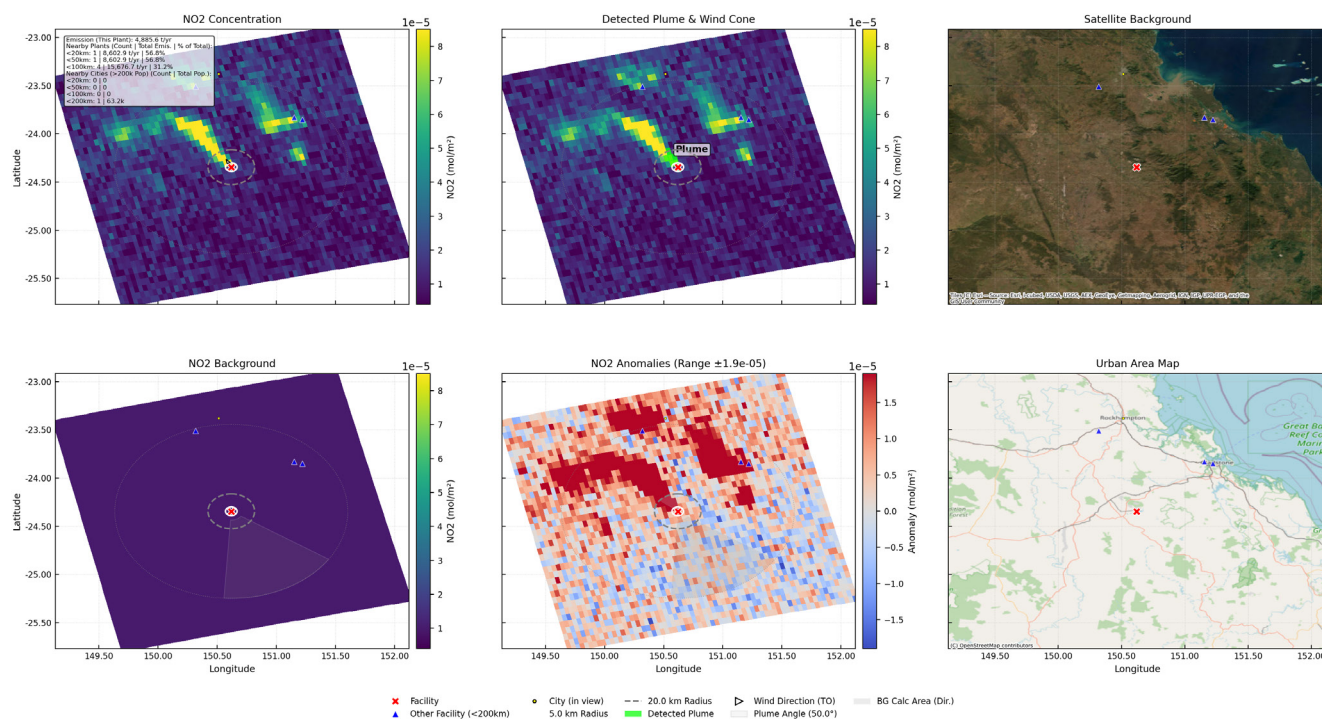


Figure F1. A clear and well-defined plume is correctly detected and attributed to the target power plant. The signal is strong and the wind direction aligns well with the plume's trajectory. Satellite Image Source: Esri, i-cubed, USDA, USGS, AEX, GeoEye, Getmapping, Aerogrid, IGN, IGP, UPR-EGP, and the GIS User Community | Powered by Esri.



Facility: CoCO2_00341 (AUS) | 2018-09-13T03:28:04.115000Z
 Annual NOx: 963.2 t/yr | Wind: 221.9° @ 1.83 m/s | Plume Detected (Area 60.0 km², Thresh 1.49e-05)

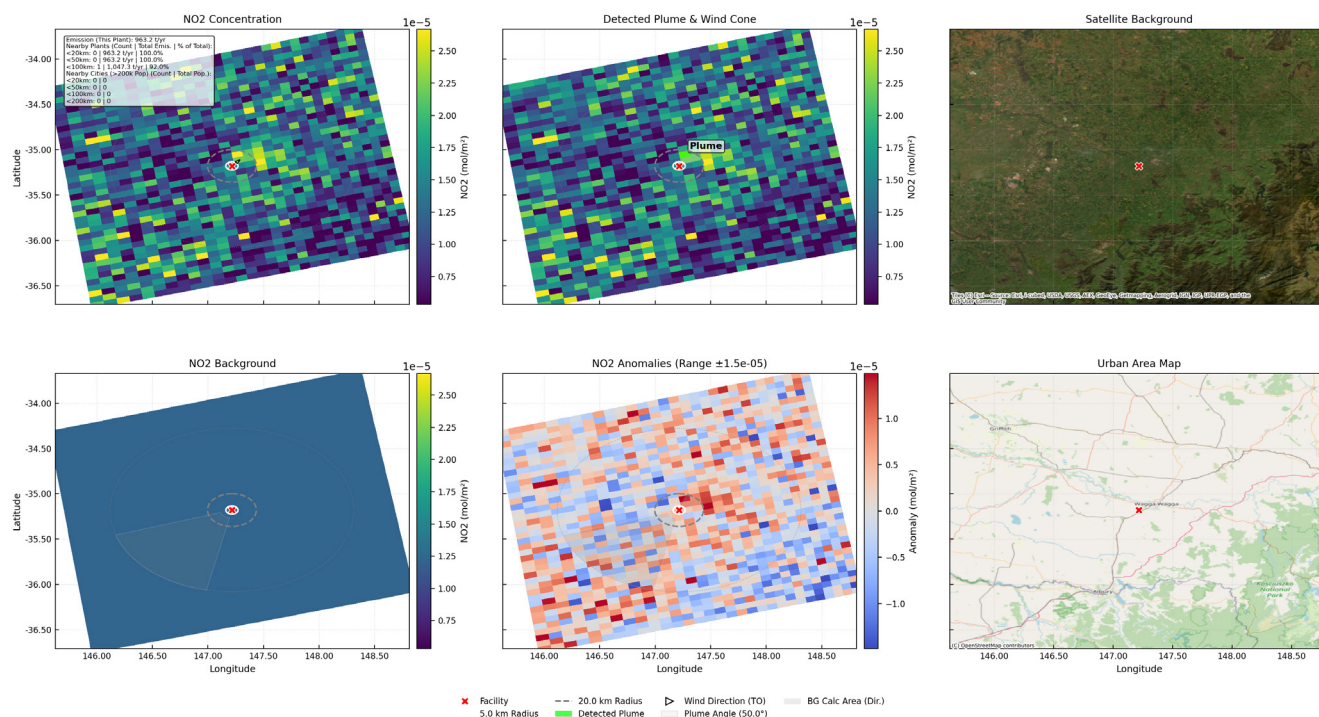


Figure F2. A successful detection of a lower-intensity plume. Despite being less concentrated, the plume's signal is distinct enough from the background to be correctly identified by the algorithm. Satellite Image Source: Esri, i-cubed, USDA, USGS, AEX, GeoEye, Getmapping, Aerogrid, IGN, IGP, UPR-EGP, and the GIS User Community | Powered by Esri.

625 F1.2 True Negatives: Correctly Identified Absence of Plumes

A true negative is recorded when no plume is detected, and none is visually present. This outcome is crucial for avoiding the false attribution of emissions on days when the plant is offline, emissions are too low, or high winds disperse the plume too rapidly to form a coherent signal.



Facility: CoCO2_00042 (ARG) | 2018-11-21T17:08:50.729000Z
Annual NOx: 418.0 t/yr | Wind: 48.9° @ 1.61 m/s | No Significant Plume (Area 0.0 km², Thresh 1.46e-05)

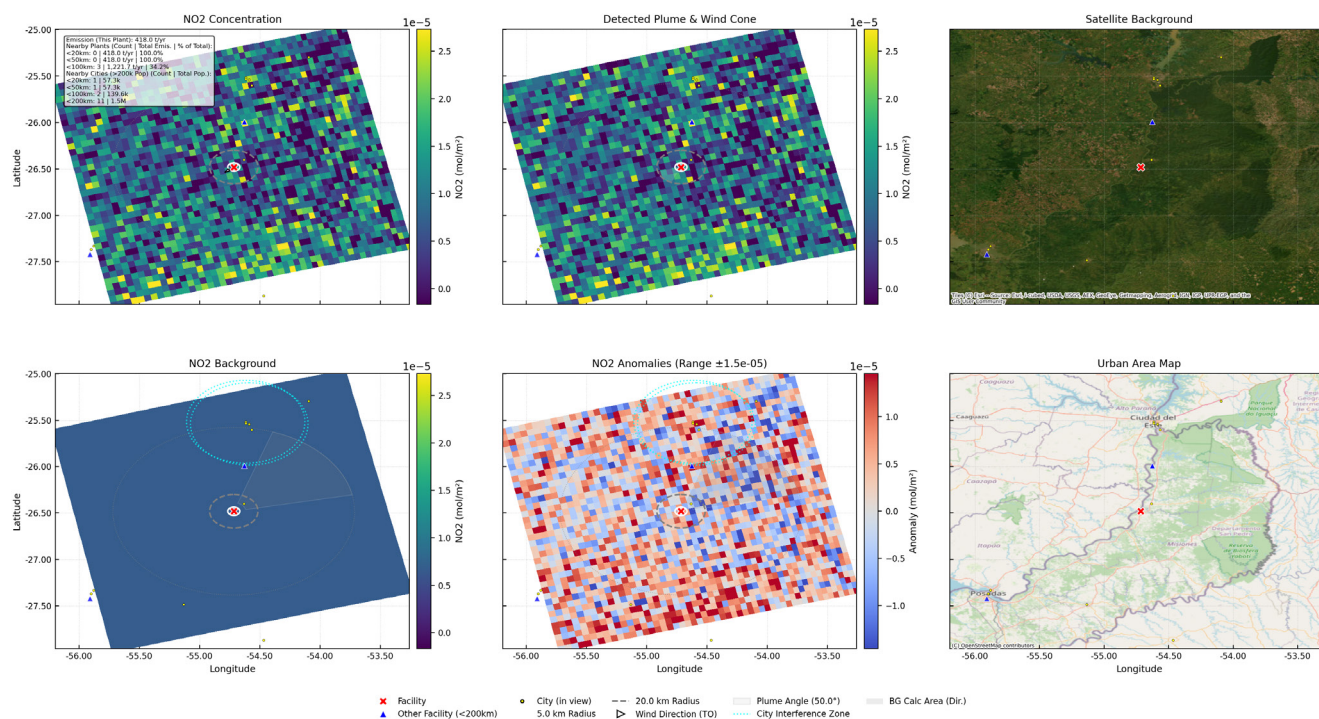


Figure F3. A correct true negative. The NO₂ map shows no discernible plume emanating from the facility, and the algorithm correctly returns no detection. Satellite Image Source: Esri, i-cubed, USDA, USGS, AEX, GeoEye, Getmapping, Aerogrid, IGN, IGP, UPR-EGP, and the GIS User Community | Powered by Esri.



Facility: CoCO2_07946 (JPN) | 2018-10-03T04:12:44.448000Z
Annual NOx: 344.7 t/yr | Wind: 16.4° @ 3.16 m/s | No Significant Plume (Area 0.0 km², Thresh 1.50e-04) | BG Mode: directional

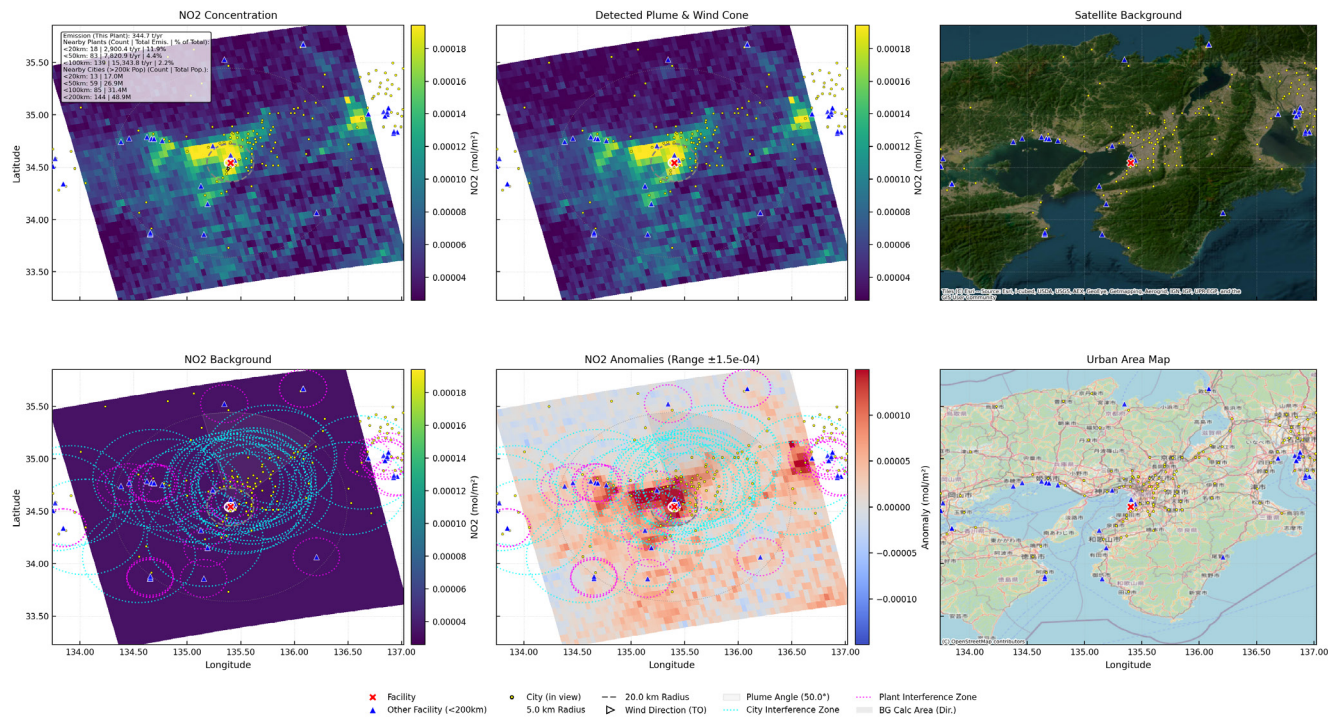


Figure F4. A correct true negative. Although a plume is visible, most likely it originates from other emission sources, and the interference zone correctly prevents it from being attributed to the power plant under study. Satellite Image Source: Esri, i-cubed, USDA, USGS, AEX, GeoEye, Getmapping, Aerogrid, IGN, IGP, UPR-EGP, and the GIS User Community | Powered by Esri.

F2 Failed Detections and Their Causes

630 F2.1 Failure Due to Interference

Interference is a primary cause of misclassification. A false positive can occur when emissions from a nearby source drift into the analysis area and are wrongly attributed to the target plant. Conversely, a false negative can happen if the algorithm's *interference mask*—designed to prevent such errors—is too aggressive and obscures a genuine plume from the target plant.



Facility: CoCO2_02070 (CHL) | 2018-09-07T18:56:42.594000Z
Annual NOx: 447.1 t/yr | Wind: 295.9° @ 3.00 m/s | Plume Detected (Area 52.9 km², Thresh 1.68e-05)

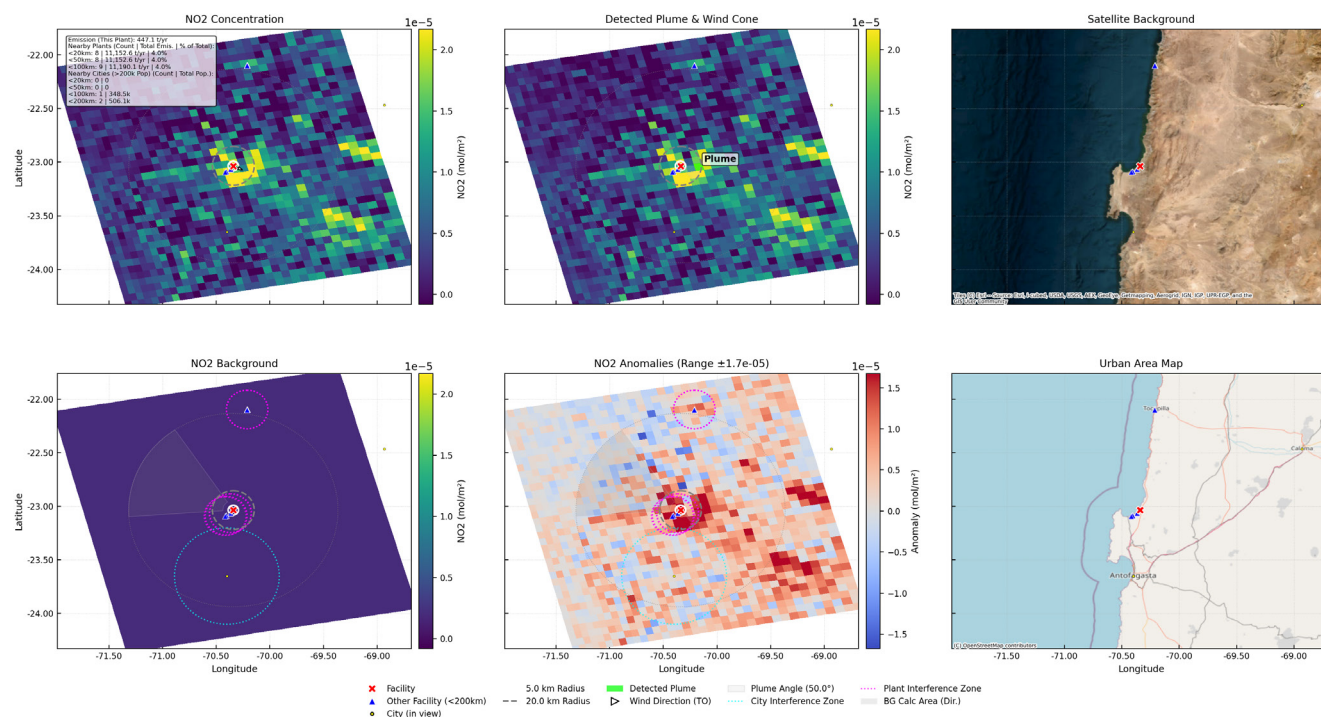


Figure F5. Cause: Interference (False Positive). A plume signal is detected, but it originates from an adjacent power plant. The algorithm incorrectly attributes this interfering emission to the target facility. Satellite Image Source: Esri, i-cubed, USDA, USGS, AEX, GeoEye, Getmapping, Aerogrid, IGN, IGP, UPR-EGP, and the GIS User Community | Powered by Esri.



Facility: CoCO2_04813 (CHN) | 2018-11-26T05:41:51.936000Z
Annual NOx: 262.6 t/yr | Wind: 210.1° @ 4.17 m/s | Plume Detected (Area 49.3 km², Thresh 4.95e-05)

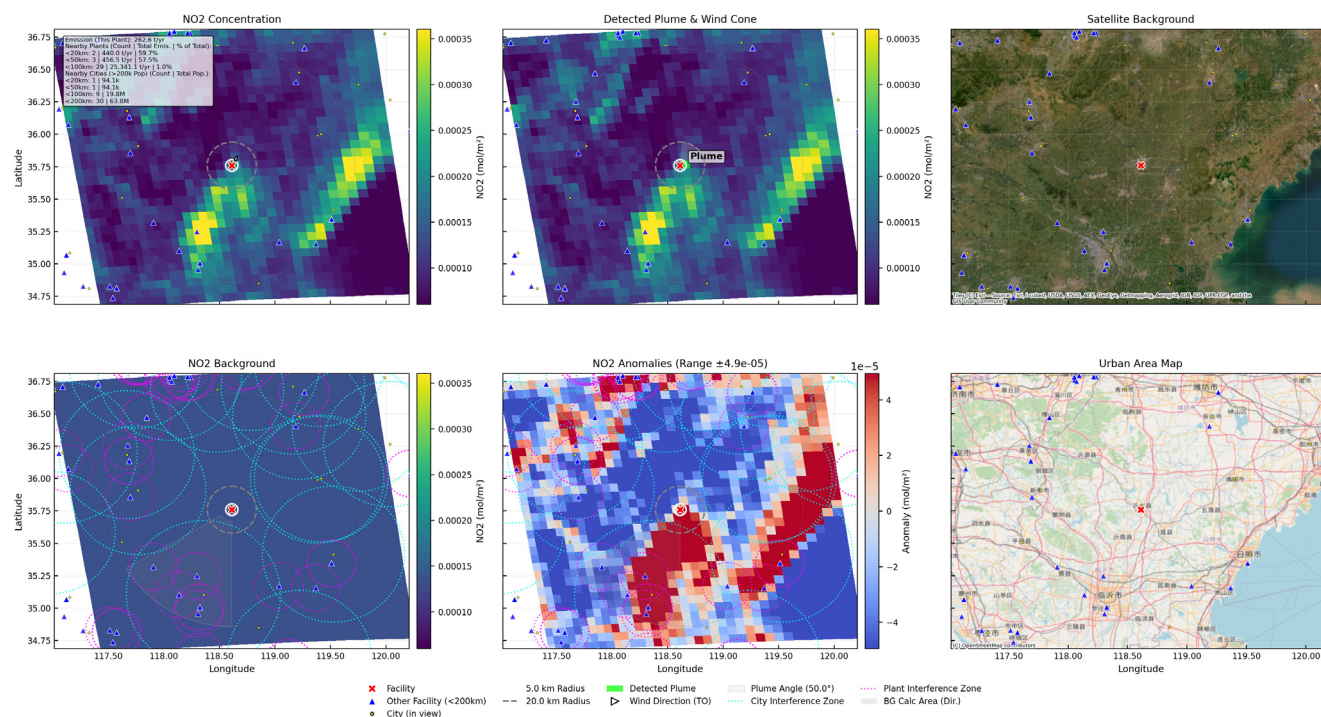


Figure F6. Cause: Interference (False Positive). Elevated NO₂ levels from a nearby urban area are mistaken for a power plant plume. This type of error is common when winds blow from a city toward the facility. Satellite Image Source: Esri, i-cubed, USDA, USGS, AEX, GeoEye, Getmapping, Aerogrid, IGN, IGP, UPR-EGP, and the GIS User Community | Powered by Esri.

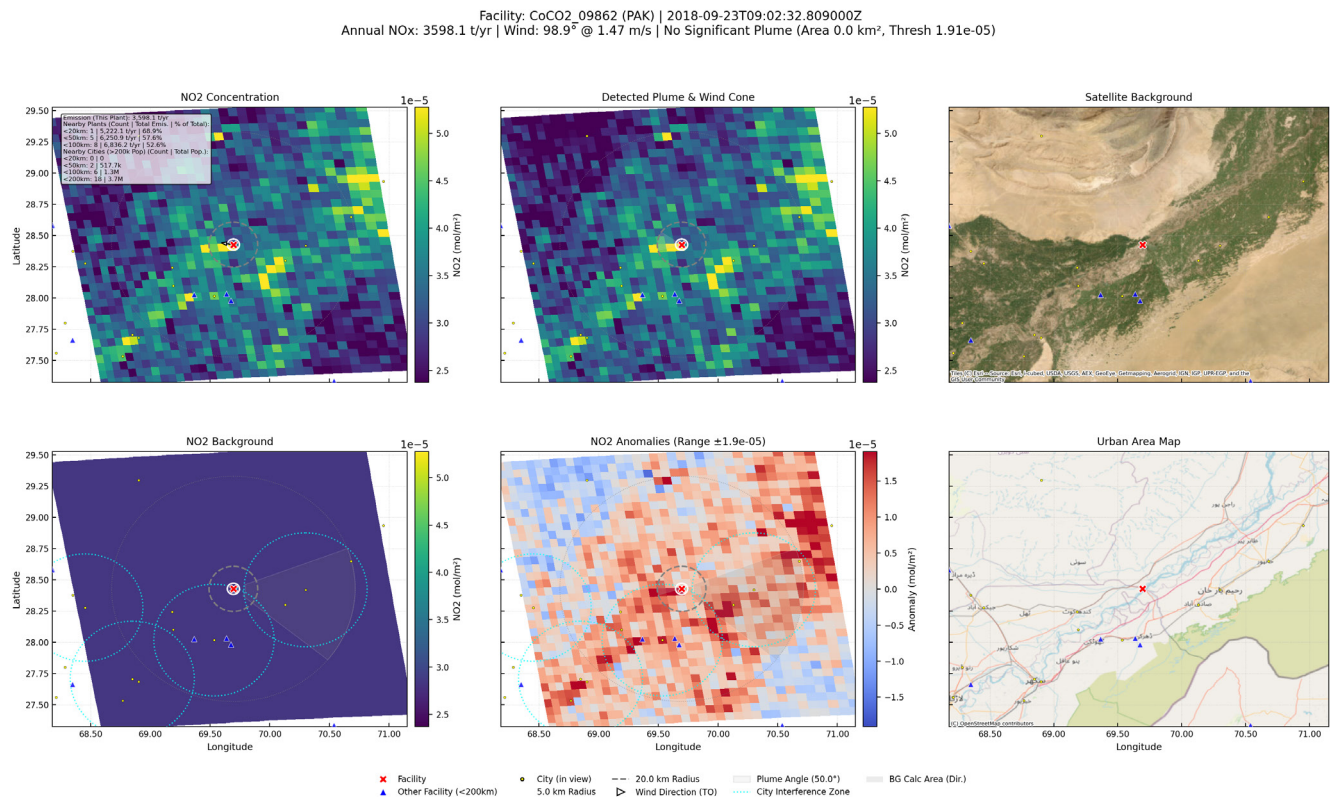


Figure F7. Cause: Interference Mask (False Negative). A prominent plume from the target plant is visible, but it is not detected because it overlaps with the pre-defined interference mask, which suppresses signals in that area to avoid confusion with other sources. Satellite Image Source: Esri, i-cubed, USDA, USGS, AEX, GeoEye, Getmapping, Aerogrid, IGN, IGP, UPR-EGP, and the GIS User Community | Powered by Esri.

F2.2 Failure Due to Incorrect Detection Threshold

635 The detection outcome is highly sensitive to the chosen NO₂ concentration threshold. If the threshold is too high, genuine but weak plumes are missed (false negative). If it is set too low, random background noise or minor, insignificant fluctuations can be flagged as a plume (false positive).



Facility: CoCO2_00629 (BRA) | 2018-09-22T17:35:53.425000Z
Annual NOx: 232.7 t/yr | Wind: 109.0° @ 1.91 m/s | Plume Detected (Area 49.3 km², Thresh 2.10e-05)

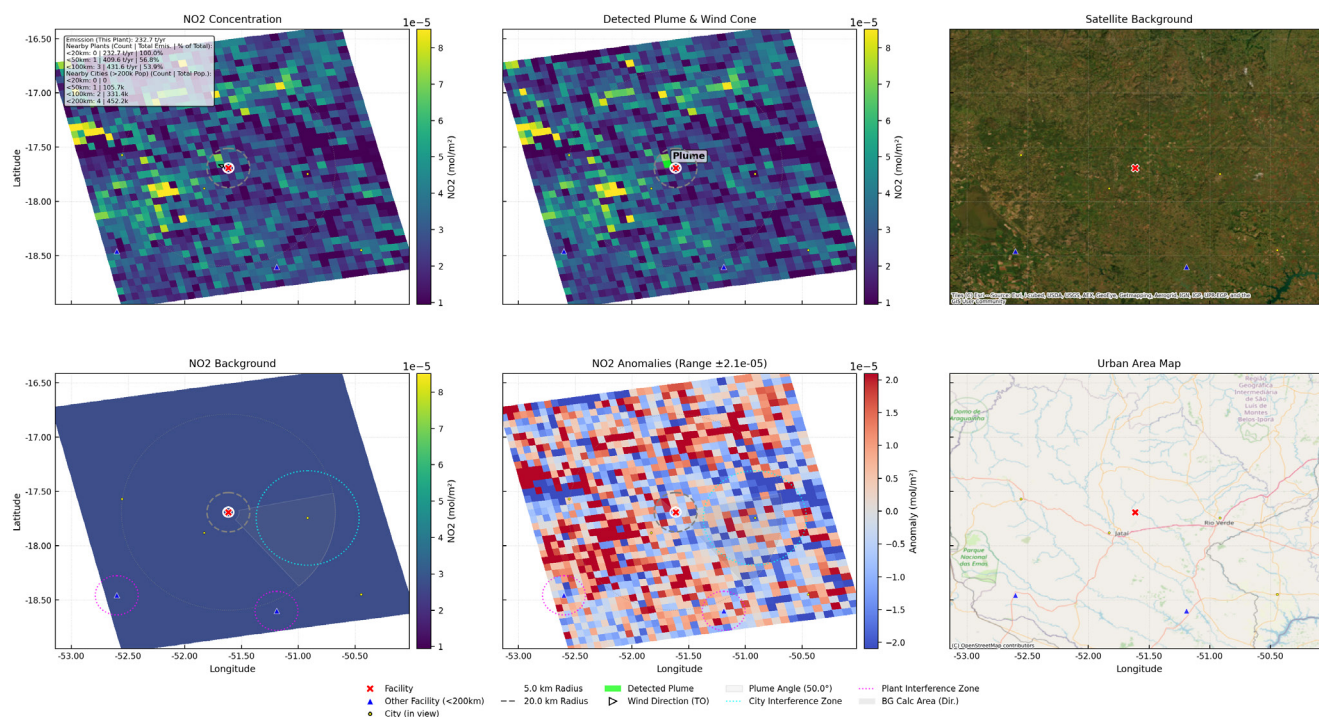


Figure F9. Cause: Low Threshold (False Positive). The detection threshold is set too low, causing the algorithm to identify a plume where none exists. The flagged area is simply part of the natural background variability in NO₂ concentration. Satellite Image Source: Esri, i-cubed, USDA, USGS, AEX, GeoEye, Getmapping, Aerogrid, IGN, IGP, UPR-EGP, and the GIS User Community | Powered by Esri.

F2.3 Failure Due to Regional Gradients

Sometimes, large-scale variations in NO₂ can create sharp gradients that the algorithm misinterprets as a plume. A classic
640 example is the land-sea boundary, where NO₂ levels are often naturally higher over terrestrial areas than over water.



Facility: CoCO2_15320 (IRL) | 2018-05-11T13:04:26.218000Z
Annual NOx: 1312.8 t/yr | Wind: 239.7° @ 6.78 m/s | Plume Detected (Area 50.9 km², Thresh 1.94e-05)

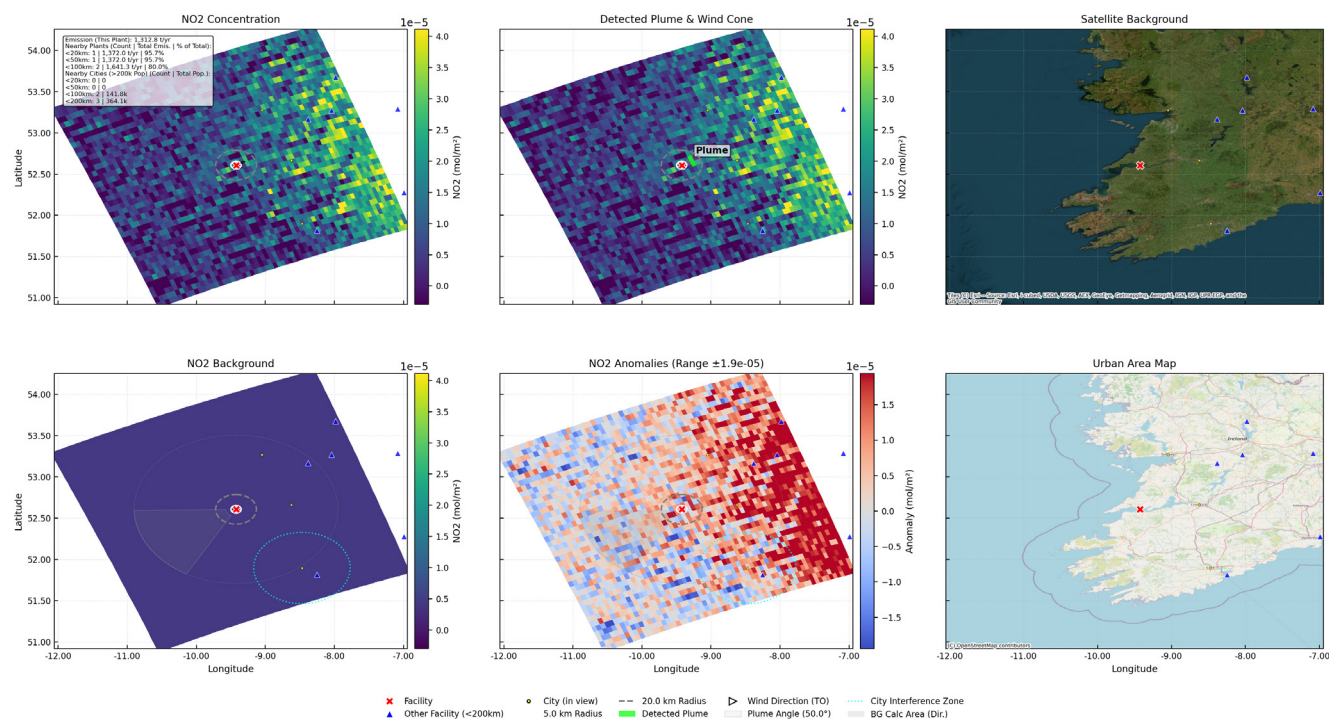


Figure F10. Cause: Topographical Gradient (False Positive). No distinct plume is present. The algorithm has incorrectly identified the sharp contrast between higher NO₂ concentrations banked against the mountain and lower concentrations in the surrounding area as an emission plume. Satellite Image Source: Esri, i-cubed, USDA, USGS, AEX, GeoEye, Getmapping, Aerogrid, IGN, IGP, UPR-EGP, and the GIS User Community | Powered by Esri.



Facility: CoCO2_00346 (AUS) | 2018-07-14T05:56:24.709000Z
Annual NOx: 230.7 t/yr | Wind: 54.0° @ 1.78 m/s | Plume Detected (Area 38.1 km², Thresh 9.31e-06)

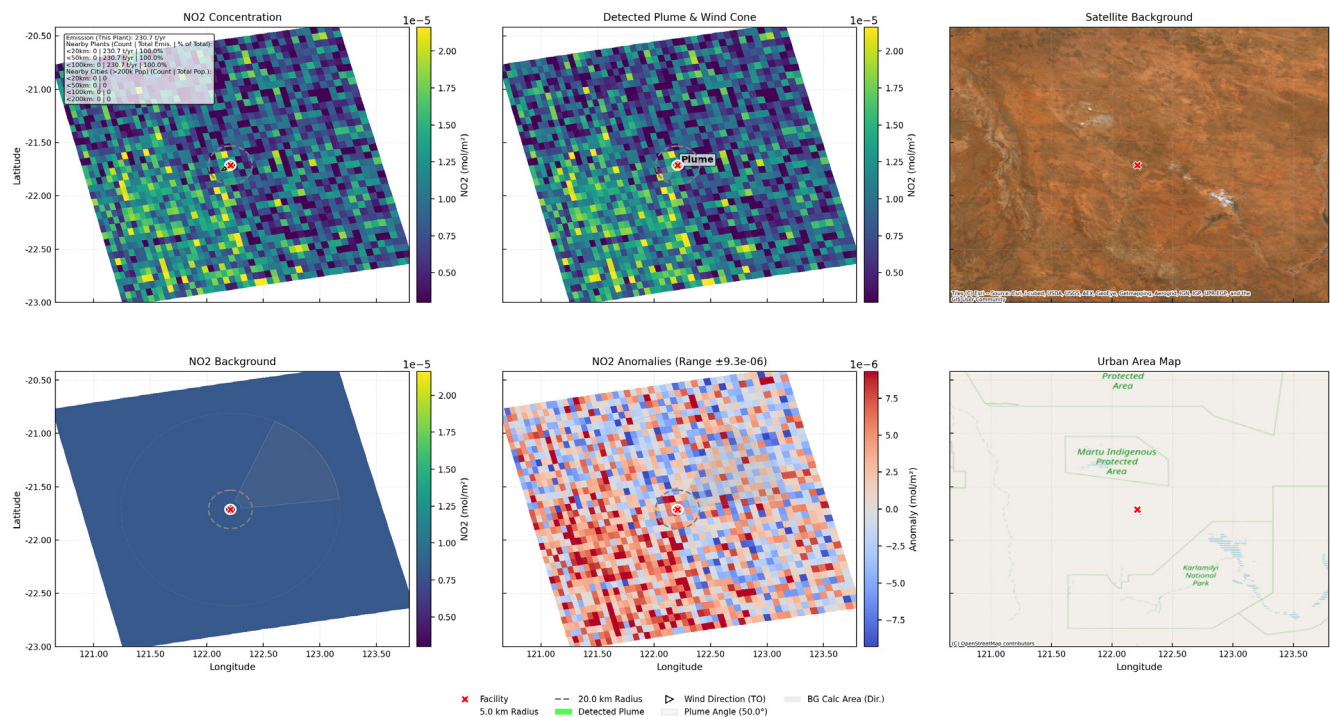


Figure F11. Cause: Gradient (False Positive). No distinct plume is present. The algorithm has incorrectly identified the sharp contrast between higher NO₂ concentrations in front of the mountain as plumes. Satellite Image Source: Esri, i-cubed, USDA, USGS, AEX, GeoEye, Getmapping, Aerogrid, IGN, IGP, UPR-EGP, and the GIS User Community | Powered by Esri.

Author contributions. RH finished the experiment and write the paper except discussion and conclusion. SW supervised the project and write the discussion and conclusion in the paper.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. We sincerely thank Hannah Grauer, Yuhao Nie, Dan Klugen for the various discussions. We thank Wenjie Lu, Wengong Jin, Ce Liu, Zituo Chen, and He Xu for their support during the project. We thank all the data providers at NASA, U.S. EPA, and ECMWF; without their long-term data programs, this study could not have been possible.



References

- CAMD's Power Sector Emissions Data Guide, office of Atmospheric Programs, Clean Air Markets Division.
- Agreement, P.: Agreement adopted on 12 December 2015 at the twenty-first session of the Conference of the Parties to the United Nations Framework Convention on Climate Change, United Nations Treaty Collection, 2015.
- 650 Beirle, S., Borger, C., Dörner, S., Li, A., Hu, Z., Liu, F., Wang, Y., and Wagner, T.: Pinpointing nitrogen oxide emissions from space, *Science advances*, 5, eaax9800, 2019.
- Bruno, J. H., Jervis, D., Varon, D. J., and Jacob, D. J.: U-Plume: automated algorithm for plume detection and source quantification by satellite point-source imagers, *Atmospheric Measurement Techniques*, 17, 2625–2636, 2024.
- 655 Couture, H. D., Alvara, M., Freeman, J., Davitt, A., Koenig, H., Rouzbeh Kargar, A., O'Connor, J., Söldner-Rembold, I., Ferreira, A., Jeyaratnam, J., et al.: Estimating carbon dioxide emissions from power plant water vapor plumes using satellite imagery and machine learning, *Remote Sensing*, 16, 1290, 2024.
- Cusworth, D. H., Duren, R. M., Thorpe, A. K., Eastwood, M. L., Green, R. O., Dennison, P. E., Frankenberg, C., Heckler, J. W., Asner, G. P., and Miller, C. E.: Quantifying global power plant carbon dioxide emissions with imaging spectroscopy, *AGU Advances*, 2, e2020AV000 350, 2021.
- 660 Cusworth, D. H., Thorpe, A. K., Miller, C. E., Ayasse, A. K., Jiorle, R., Duren, R. M., Nassar, R., Mastrogiacono, J.-P., and Nelson, R. R.: Two years of satellite-based carbon dioxide emission quantification at the world's largest coal-fired power plants, *Atmospheric Chemistry and Physics*, 23, 14 577–14 591, 2023.
- Ding, N., Xi, Y., Jiang, W., Li, H., Su, J., Yang, S., and Lie, T. T.: State-of-the-art carbon metering: Continuous emission monitoring systems for industrial applications, *Heliyon*, 11, 2025.
- 665 Dumont Le Brazidec, J., Vanderbecken, P., Farchi, A., Broquet, G., Kuhlmann, G., and Bocquet, M.: Quantification of CO₂ hotspot emissions from OCO-3 SAM CO₂ satellite images using deep learning methods, *Geoscientific Model Development*, 18, 3607–3622, 2025.
- Eskes, H. and Boersma, K.: Averaging kernels for DOAS total-column satellite retrievals, *Atmospheric Chemistry and Physics*, 3, 1285–1291, 2003.
- 670 European Commission: Industrial Emissions Directive (2010/75/EU), <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32010L0075>, 2010.
- European Environment Agency: Reported data on large combustion plants covered by the Industrial Emissions Directive (2010/75/EU) (version 5.2), <https://www.eea.europa.eu/data-and-maps/data/lcp-9>, accessed 2025-10-07, 2019.
- European Environment Agency: The European Pollutant Release and Transfer Register (E-PRTR), Member States reporting under Article 7 of Regulation (EC) No 166/2006 (version 18), <https://www.eea.europa.eu/data-and-maps/data/member-states-reporting-art-7-under-the-european-pollutant-release-and-transfer-register-e-prtr-regulation-23>, accessed 2025-10-07, 2020.
- 675 Goldberg, D. L., Lu, Z., Streets, D. G., de Foy, B., Griffin, D., McLinden, C. A., Lamsal, L. N., Krotkov, N. A., and Eskes, H.: Enhanced capabilities of TROPOMI NO₂: Estimating NO_x from north american cities and power plants, *Environmental science & technology*, 53, 12 594–12 601, 2019.
- 680 Guevara, M., Enciso, S., Tena, C., Jorba, O., Dellaert, S., Denier van der Gon, H., and Pérez García-Pando, C.: A global catalogue of CO₂ emissions and co-emitted species from power plants, including high-resolution vertical and temporal profiles, *Earth System Science Data*, 16, 337–373, <https://doi.org/10.5194/essd-16-337-2024>, 2024.



- Guo, W., Shi, Y., Liu, Y., and Su, M.: CO₂ emissions retrieval from coal-fired power plants based on OCO-2/3 satellite observations and a
685 Gaussian plume model, *Journal of Cleaner Production*, 397, 136 525, 2023.
- Kuhlmann, G., Koene, E., Meier, S., Santaren, D., Broquet, G., Chevallier, F., Hakkarainen, J., Nurmela, J., Amorós, L., Tamminen, J., et al.:
The ddeq Python library for point source quantification from remote sensing images (version 1.0), *Geoscientific Model Development*, 17,
4773–4789, 2024.
- Liu, F., Duncan, B. N., Krotkov, N. A., Lamsal, L. N., Beirle, S., Griffin, D., McLinden, C. A., Goldberg, D. L., and Lu, Z.: A methodology to
690 constrain carbon dioxide emissions from coal-fired power plants using satellite observations of co-emitted nitrogen dioxide, *Atmospheric
Chemistry and Physics*, 20, 99–116, 2020.
- M, C., D, G., M, B., E, S., M, M., E, S., F, P., F, M.-F., J, O., R, Q., A, R. M., P, T.-M., G, G., S, R., D, J. F. O., A, B., J, S.-M.-A., and
E, V.: CO₂ emissions of all world countries, Scientific analysis or review KJ-NA-31-182-EN-N (online), KJ-NA-31-182-EN-C (print),
Luxembourg (Luxembourg), ISBN 978-92-76-55802-6 (online), 978-92-76-55803-3 (print), ISSN 1831-9424 (online), 1018-5593 (print),
695 <https://doi.org/10.2760/07904> (online), 10.2760/730164 (print), 2022.
- McDuffie, E. E., Smith, S. J., O'Rourke, P., Tibrewal, K., Venkataraman, C., Marais, E. A., Zheng, B., Crippa, M., Brauer, M., and Mar-
tin, R. V.: A global anthropogenic emission inventory of atmospheric pollutants from sector- and fuel-specific sources (1970–2017): an
application of the Community Emissions Data System (CEDS), *Earth System Science Data*, 12, 3413–3442, [https://doi.org/10.5194/essd-
12-3413-2020](https://doi.org/10.5194/essd-12-3413-2020), 2020.
- 700 Palmer, P. I., Jacob, D. J., Chance, K., Martin, R. V., Spurr, R. J., Kurosu, T. P., Bey, I., Yantosca, R., Fiore, A., and Li, Q.: Air mass factor
formulation for spectroscopic measurements from satellites: Application to formaldehyde retrievals from the Global Ozone Monitoring
Experiment, *Journal of Geophysical Research: Atmospheres*, 106, 14 539–14 550, 2001.
- Rouet-Leduc, B. and Hulbert, C.: Automatic detection of methane emissions in multispectral satellite imagery using a vision transformer,
Nature Communications, 15, 3801, 2024.
- 705 SimpleMaps: World Cities Database, <https://simplemaps.com/data/world-cities>, [accessed June 18, 2025], n.d.
- Srivastava, R. P., Kumar, S., and Tiwari, A.: Continuous emission monitoring systems (CEMS) in India: performance evaluation, policy gaps
and financial implications for effective air pollution control, *Journal of Environmental Management*, 359, 120 584, 2024.
- U.S. Environmental Protection Agency: Clean Air Act, <https://www.epa.gov/clean-air-act-overview>, 1970.
- U.S. Environmental Protection Agency: Emissions & Generation Resource Integrated Database (eGRID), 2018, <https://www.epa.gov/egrid>,
710 data year: 2018; Published: 2020. Accessed: 2025-05-19, 2020.
- U.S. Environmental Protection Agency: Progress Report on Emission Controls and Monitoring, [https://www.epa.gov/power-sector/
progress-report-emission-controls-and-monitoring](https://www.epa.gov/power-sector/progress-report-emission-controls-and-monitoring), accessed: 2025-06-15, 2023.
- Van Geffen, J., Boersma, K. F., Eskes, H., Sneep, M., Ter Linden, M., Zara, M., and Veefkind, J. P.: S5P TROPOMI NO₂ slant column
retrieval: Method, stability, uncertainties and comparisons with OMI, *Atmospheric Measurement Techniques*, 13, 1315–1335, 2020.