

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~~detectability from space. Using ~~TROPOMI observations~~TROPospheric Monitoring Instrument (TROPOMI) observations (nadir pixel size 3.5–7 km) over 6,000 of the world’s highest-emitting power plants and hourly ~~CEMS~~Continuous Emissions Monitoring Systems (CEMS) data for 500 U.S. plants, we develop an automated algorithm that labels plumes and attributes them to their sources with 98% accuracy. For the subsequent detectability analysis, we restrict to plants outside interference zones (at least 20 km from other major power plants and 45–90 km from cities (depending on city size)), which retains 45.0% of U.S. and 21.1% of global NO_x emissions in our datasets. We then train a machine learning model to predict plume detectability ~~from environmental~~(the probability of detection given the observation conditions) from meteorological, environmental, sensor, and power-plant variables sampled at the single TROPOMI pixel over each plant (F1 score > 0.66, 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 altitude~~, surface albedo (NO₂ window), sensor zenith angle, primary fuel type, and wind speed jointly explain much of the ~~detection variability~~. An variability in detectability. For U.S. power plants, an hourly NO_x emission rate of ≈ 400 kg/h~~corresponds to a~~h⁻¹ corresponds to $\sim 50\%$ detection probability on average, but detection rates vary detectability, but detectability varies 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~~TROPOMI 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~~NO_x and over 40% of ~~CO₂~~CO₂ emissions globally (McDuffie et al., 2020; Crippa et al., 2022). Accurate quantification of these emissions is essential for enforcing ~~environmental regulations and achieving air-quality regulations~~regulations (U.S. Environmental Protection Agency, 1970; European Commission, 2010) and tracking progress toward international climate goals (~~Paris Agreement, 2015; U.S. Environmental Protection Agency, 1970; European Commission, 2010~~)(Paris Agreement). 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 Emissions Monitoring Systems (CEMS)—~~which, which~~ measure stack-level emissions ~~in real time continuously~~ using instruments like ~~NO_x NO_x~~ analyzers, flow meters, and opacity ~~monitors—enable monitors and report hourly averages, enable~~ direct, high-frequency reporting (Ding et al., 2025). In contrast, most power plants in low- and middle-income countries lack such
30 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 ~~NO_x~~ 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)~~
35 ~~Early demonstrations with the Ozone Monitoring Instrument (OMI) established wind-rotated downwind-plume methods for extracting emission rate and lifetime, and applied them to tens of major cities and power plants (Beirle et al., 2011; Liu et al., 2016; de Foy et al., 2016). TROPOMI’s finer spatial resolution has substantially advanced this line of work: Exponentially-Modified Gaussian (EMG) and related top-down methods have been extended to power plants in North America and beyond (Goldberg et al., 2019; Lange et al., 2022; Beirle et al., 2022), and the divergence method has produced global point-source catalogs that now contain over a thousand identified NO_x sources (Beirle et al., 2019, 2021, 2023).~~ 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-Ledue and Hulbert, 2024; Cusworth et al., 2023; Dumont Le Brazidec et al., 2025).~~

A complementary line of research has focused on plume identification rather than quantification, asking whether a plume is discernible on a given overpass and using automated or learning-based methods to scale detection across many sources.
45 ~~Supervised image classification has been demonstrated for NO₂ on TROPOMI (Finch et al., 2022), while the analogous methane and carbon dioxide literature is more developed, with detection and quantification frameworks for individual plumes and super-emitters across multiple sensors (Kuhlmann et al., 2019; Cusworth et al., 2021, 2023; Lauvaux et al., 2022; Schuit et al., 2023; Rouet-Ledue et al., 2024). Theoretical expectations for plume column behavior are well established: for a steady point source, the downwind column enhancement scales linearly with the emission rate and inversely with the product of wind speed and the horizontal spread of the plume (Varon et al., 2018), while the effective NO_x lifetime is modulated by photolysis, the OH concentration, and temperature (Valin et al., 2013; Romer et al., 2018; Laughner and Cohen, 2019). On the retrieval side, sensitivity depends on air-mass factor (AMF) geometry, surface and cloud conditions, and background noise (Palmer et al., 2001; Boersma et al., 2011; Eskes and van den Broek, 2009). To date, no published work has systematically studied power plant plume detectability at the global scale, nor modeled when and why plumes become observable in TROPOMI data.~~

55 Amidst the current satellite-based efforts to monitor anthropogenic ~~NO₂ NO_x~~ emissions, a challenge persists: power plant plumes with similar NO_x emissions exhibit vastly different detectability ~~across regions (Figure 1), (the probability of detection given the observation conditions) across regions, as illustrated in Figure 1. This variability is~~ driven by meteorology, ~~surface albedo (NO₂ window),~~ 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 detec-
60 tion 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 implications for improving satellite-based $\text{NO}_x\text{-NO}_x$ inverse modeling as well as future sensor design.

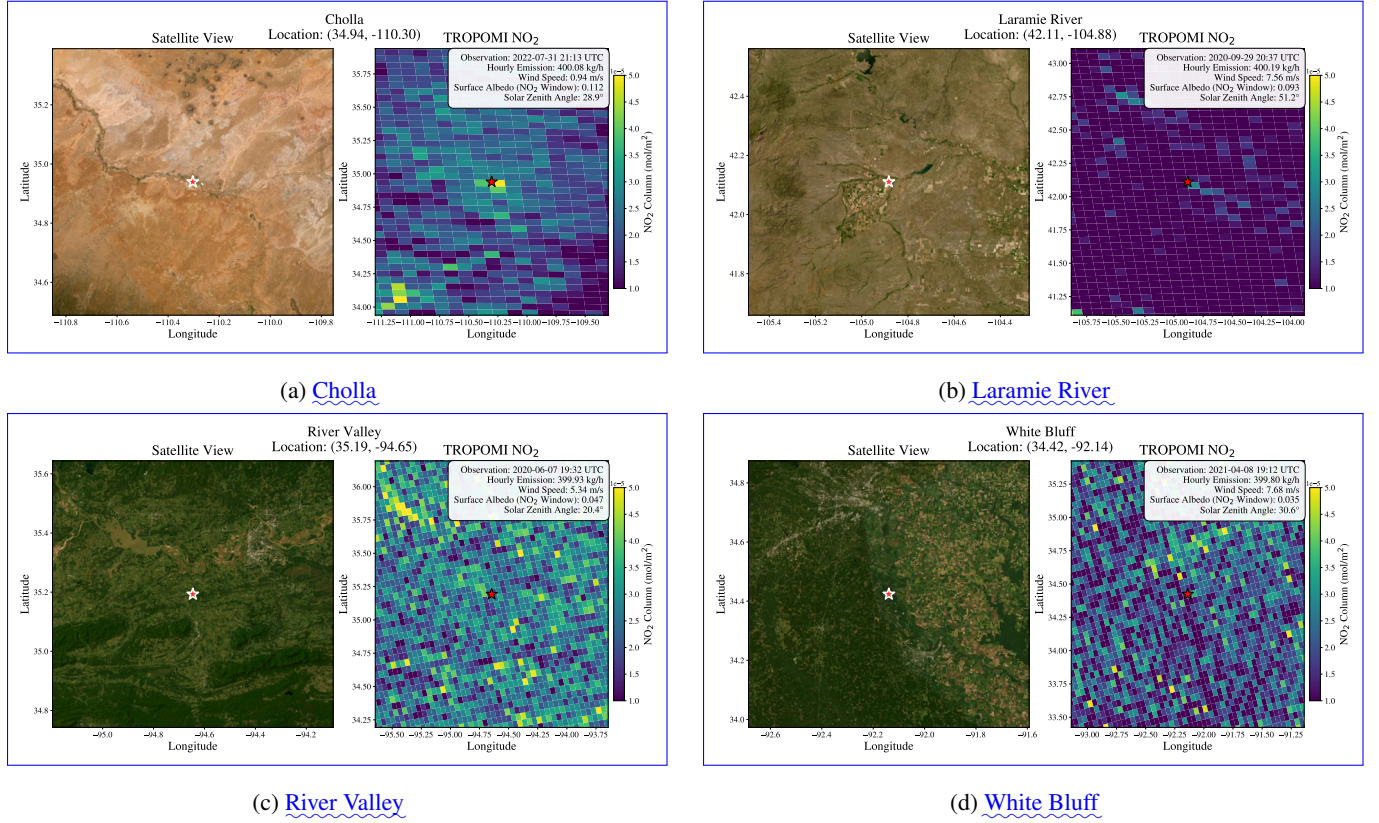


Figure 1. Comparison of tropospheric NO_2 vertical column densities observed by satellite over four U.S. power plants with closely matched hourly emissions ($\approx 400 \text{ kg h}^{-1}$). NO_2 plumes are visible near Cholla and Laramie River (top row); no significant enhancement is detected near River Valley or White Bluff (bottom row). The TROPOMI overpass time and observed conditions (100 m wind speed, surface albedo in the NO_2 window, solar zenith angle) are annotated on each panel. Basemap: ArcGIS World Imagery (Esri, n.d.) via contextily (Arribas-Bel and contextily contributors, 2024).

Specifically, we aim to systematically explain the spatiotemporal variability of $\text{NO}_2\text{-NO}_2$ plume detectability from power plants using satellite observations. To do so, we develop a machine learning pipeline that first labels $\text{NO}_2\text{-NO}_2$ plumes in TROPOMI (Sentinel-5P) images and then predicts plume detectability using features including meteorology, sensor geometry, environment variables, and meteorology, and power-plant 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 $\text{NO}_x\text{-NO}_x$ inventory.

Our contributions are as follows.

1. We present a plume-labeling-plume-detection algorithm that achieves $\sim 98\%$ accuracy against manual annotations, designed for point source NO_2 attribution.
2. We produce the first global maps of plume detectability across 6more than 1,000 power plants worldwide, revealing strong geographic patterns. We For U.S. power plants, we find that an average hourly emission rate of ≈ 400 kg/h corresponds to a h^{-1} corresponds to $\sim 50\%$ detection probability on averaged detectability, although detectability varies widely depending on location, sensing conditions, and meteorology.
3. We quantify how meteorological, environmental, and observational features influence NO_2 sensor, and power-plant features influence NO_2 plume detectability across both regional and global scales. Using Permutation Score and SHAP SHapley Additive exPlanations (SHAP) Score Estimation, we identify NO_x Emission, Surface Albedo, Altitude, Surface Albedo (NO_2 Window), Sensor Zenith Angle, Primary Fuel Type, 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~~ Together, these results clarify 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 and provide~~ empirical relationships that can inform future satellite-based emission quantification.

~~Jim Bridger Bonanza Watson Electric Generating Plant Gerald Gentleman Station Comparison of tropospheric NO_2 vertical column densities observed by satellite over four U.S. power plants. Although all facilities reported the same hourly emissions (400 kg/h), NO_2 plumes are visible only near the Jim Bridger and Bonanza (top row). In contrast, no significant NO_2 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~~ Our analysis combines five input data sources: Sentinel-5P TROPOMI Level-2 NO_2 retrievals (Veefkind et al., 2012; Van Geffen et al., 2020), which provide both the satellite NO_2 observations and a set of sensor- and scene-level variables (e.g., sensor zenith angle, surface albedo, cloud fraction); U.S. EPA Clean Air Markets Program Data (CAMPD) records (U.S. Environmental Protection Agency, 2023) for hourly NO_x emissions at 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

100 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).

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 [U.S. power plants](#); the [CoCO₂ power-plant emissions catalogue](#) to identify the 6,000 facilities with the largest 110 [global power-plant emission catalog](#) (Guevara et al., 2024) for 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 115 each other, then removed any in those groups that shared identical emission values across all reported pollutants (CO_2 , NO_x , SO_x , CO , and CH_4). 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. 120 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](#) at 6,000 plants worldwide; [European Centre for Medium-Range Weather Forecasts \(ECMWF\) ERA5 reanalysis](#) (Hersbach et al., 2020) for meteorological variables; and the [SimpleMaps World Cities Database](#) (SimpleMaps, n.d.) for the locations and populations of nearby urban areas used in the interference filter. The subsections below describe each source and the resulting feature set.

125 2.1 Satellite NO_2 Measurements

To form the observation dataset, we used satellite data of tropospheric NO_2 vertical column densities ([Van Geffen et al., 2020](#)) from the TROPOMI instrument aboard the [European Space Agency's \(ESA\) Copernicus Sentinel-5P satellite](#) ([Veefkind et al., 2012](#)). Launched in October 2017, TROPOMI provides data at a nadir spatial resolution of [7 km \$\times\$ 3.5 km \(across-track\) \$\times\$ 3.5 km \$\times\$ 7 km \(along-track\)](#), which improved to [5.5 km \$\times\$ 3.5 km in August 2019; 3.5 \$\times\$ 5.5 km in August 2019; pixel size increases](#) 130 [toward the swath edges due to viewing geometry](#).

We downloaded raw Level-2 NO_2 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

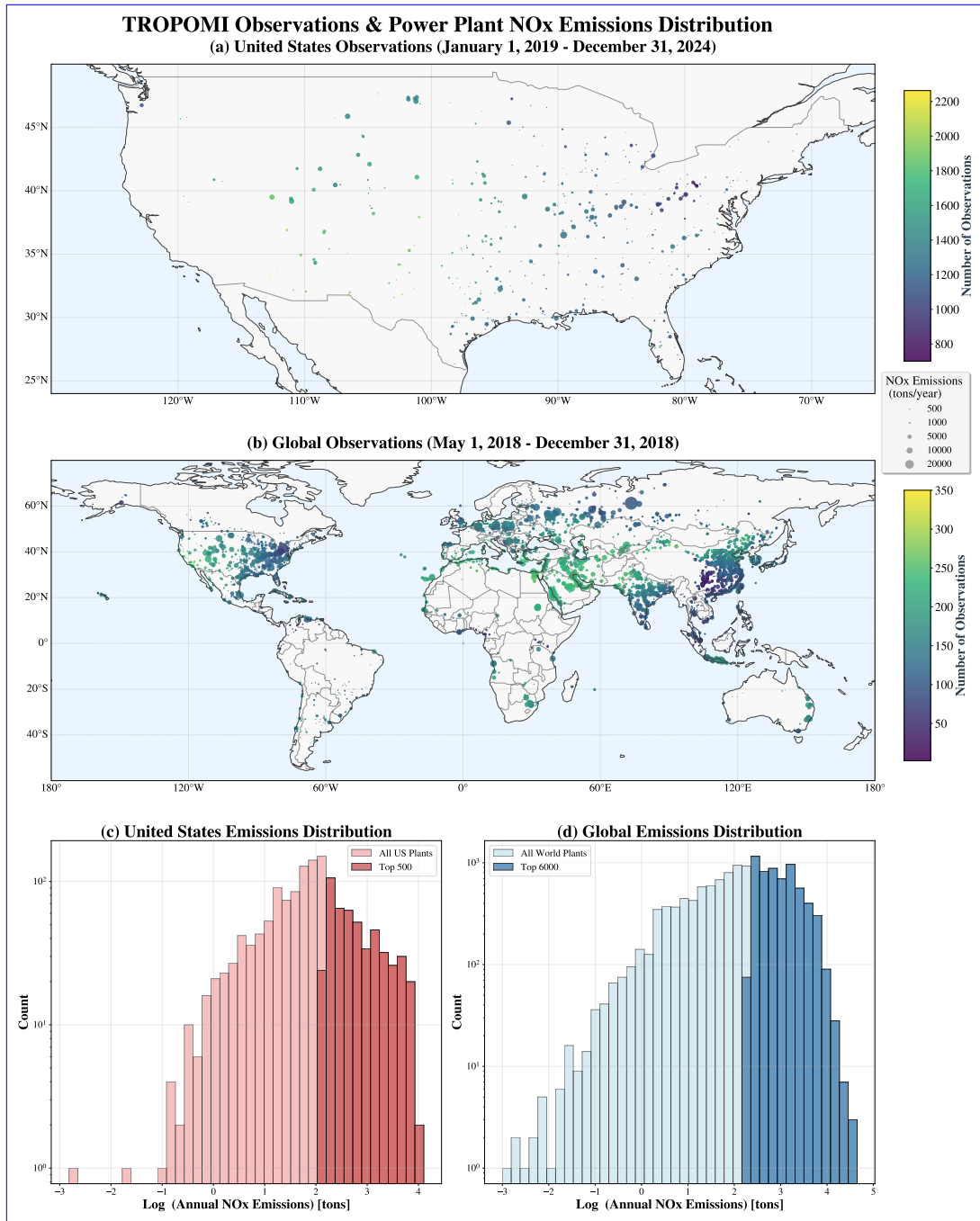


Figure 2. Distribution of TROPOMI observations and power plant NO_x emissions. Panels (a) and (b) show spatial distribution of power plant NO_x emissions (marker size) overlaid with TROPOMI observation frequency (marker color) for the United States and global facilities, respectively. Panels (c) and (d) display emission distribution histograms on a log scale showing all plants (light colors) versus analysis subsets of top emitters (dark colors) for the United States (top 500 plants) and global (top 6000 plants). TROPOMI observation periods: U.S. data from January 2019 to December 2024, global data from May to December 2018. Dataset: full quality-filtered dataset ($n = 666,222$ U.S. and $n = 875,686$ global TROPOMI observations).

1 – December 31, 2018), we obtained 3,421 files. The data come from the v2 series of the TROPOMI L2 NO₂ processor (v2.4.0–v2.8.0), spanning both the reprocessed (RPRO) and operational offline (OFFL) streams. The May 2018 start of the global window is set by the v2.4.0 NO₂ full-mission reprocessing, which begins on 1 May 2018; earlier observations are only available in the v1 product line, which is not directly comparable.

Throughout this paper we use observation to refer to one TROPOMI overpass at a target power plant: the L2 NO₂ retrieval at the pixel closest to the plant’s center, paired with the corresponding emission record.

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 ([Van Geffen et al., 2022](#)),
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~~ this process yielded a final dataset of ~~504,165~~ 666,222 high-quality observations for the U.S. analysis and ~~875,580~~ 686 for the global analysis. Across the mapped facilities (U.S., $N = 500$; global, $N = 6,000$), snapshot observation coverage is extensive. In the U.S., plants have a median of 1,301 ~~snapshots~~ ($p_{10} = 1,050.0$, $p_{90} = 1,644.7$) observations (10th and 90th percentiles $P_{10} = 1,050$ and $P_{90} = 1,645$); every plant has ≥ 700 ~~snapshots~~ (range 701–2) observations (range 701–2,263). Globally, the median is 141 ~~snapshots~~ ($p_{10} = 90$, $p_{90} = 218$) observations ($P_{10} = 90$, $P_{90} = 218$), with 98.8% of plants having ≥ 50 ~~snapshots~~ (range 2–345) observations (range 2–350). 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.

2.2 Hourly NO_x Emissions in the U.S.

~~The CAMPD, maintained by the U.S. Environmental Protection Agency’s (EPA) Clean Air Markets Program Data (CAMPD)~~ provides hourly emissions data for CO₂, NO_x, SO₂, and mercury from individual power plants. This level of granularity is unique globally and offers 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)~~ 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 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 the 500 power plants with the largest total NO_x emissions from 2019 to 2024, representing the upper tail of all U.S. facilities, for our study; their emission distribution is shown in panel (c) of Figure 2. Our selection accounts for 93.6% of the inventory’s NO_x emissions and 79.8% of its CO₂ emissions. By 2024,

of this 500-plant dataset, natural gas facilities (305 plants, 61%) and coal plants (158 plants, 32%) dominated, with smaller categories including natural gas (14), coal refuse (4), diesel oil (2), and other gas (2); plants that burn multiple fuels are counted in each applicable category. Over this six-year period, this group of top emitters transitioned measurably from coal to natural gas: 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). For context, the EPA's broader power sector data covered approximately 96% of U.S. fossil fuel-based electricity generation as of 2018 (U.S. Environmental Protection Agency, Clean Air Markets Division, 2022).

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 504,165 samples, each pairing a TROPOMI image with that hour's reported NO_x emissions.

2.3 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 reported global power plant emissions and provides annual data for CO₂, NO_x , SO_x, CO, and CH₄. The CoCO₂ catalog derived the annual emissions using a bottom-up methodology as follows:

1. **Europe:** Use each plant's official reported NO_x from EPRTTR and LCP the European Pollutant Release and Transfer Register (E-PRTR) and the Large Combustion Plants (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's Emissions & Generation Resource Integrated Database (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 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 Plume detectability is determined at the moment of each observation; relying on annual emissions therefore limits our ability to predict the plume detection per-observation detectability.

From this catalog, we selected the 6,000 plants that account facilities with the largest annual NO_x emissions in 2018, with emissions ranging from 219 to 44,548 tyr^{-1} , representing the upper tail of global power plant emissions; their emission distribution is shown in panel (d) of Figure 2. This selection accounts for 96.1% of the inventory's NO_x and 91.5% of its CO₂ emissions. This selection represents a significant portion of worldwide power plant pollution, corresponding, and corresponds

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~~
 200 ~~a TROPOMI image with the plant’s annual NO~~Coal (3,023), natural gas (1,810), and oil (690) plants dominate the set, which also includes facilities that burn biomass (354) and waste (123).

To avoid double-counting in our source-attribution analysis, we filtered the catalog for duplicate entries (defined as plants within 3 km sharing identical reported emissions across all five pollutants: CO_2 , NO_x emissions, SO_x , CO, and CH_4). This step removed 3,320 of 16,461 candidate facilities (20% of the catalog), with the bulk of removals concentrated in China and
 205 India (66% combined); per-country statistics and a brief discussion of the underlying catalog-integration causes are provided in Appendix B.

2.4 Features Predictive of Plume Detectability

~~We selected features~~ Table 1 lists the features we selected 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
 210 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.
- 215 2. ERA5 meteorological fields were ~~interpolated to plant locations using nearest-grid interpolation~~ matched to each TROPOMI overpass using nearest-neighbor interpolation in both space (to the plant location) and time (to the overpass timestamp).
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.

2.4.1 Sensor

220 Sensor zenith and azimuth angles, i.e. the viewing zenith angle (VZA) and viewing azimuth angle from the TROPOMI L2 product, 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~~ The VZA also sets the across-track ground-pixel footprint, which grows from ~ 3.5 km at nadir to up to ~ 14 km at the swath edge (Van Geffen et al., 2020), and is itself a likely driver of plume detectability. Sensor altitude varies only modestly
 225 along Sentinel-5P’s near-circular sun-synchronous orbit ($\leq 2\%$ of the mean in our data), so its direct effect on ground-pixel size and swath geometry is small; it is closely related to latitude and primarily serves as a latitude/orbit-geometry proxy alongside the explicit solar and viewing angles. 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.

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, <u>surface pressure</u> , solar zenith angle, solar azimuth angle, apparent scene pressure, aerosol index 354–388
	ERA5	<u>100 m wind speed</u> , 2 m temperature, total column water vapour, TOA <u>top-of-atmosphere (TOA)</u> 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~~ Cloud-product variables provided by the FRESKO (Fast REtrieval Scheme for Clouds from the Oxygen A-band) cloud retrieval, which models clouds as Lambertian reflectors under the Clouds-as-Reflecting-Boundaries (CRB) ~~model, a simplified Lambertian-reflector approach used by the ROCINN algorithm to retrieve cloud properties approximation.~~

2.4.2 Power plant

230 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~~detectability.

2.4.3 Meteorology

235 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~~the 100 m horizontal winds ~~wind speed derived from the ERA5 horizontal wind components~~ (u, v) ~~available from the TROPOMI product.~~ as a single feature.
- **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~~ TROPOMI 354–388 nm ~~ultraviolet (UV) aerosol index captures absorbing aerosols (e.g., dust, smoke, volcanic ash), which are an important aerosol-related interference in NO₂ retrievals; aerosol optical depth (AOD), which captures total aerosol loading (both~~ absorbing and non-absorbing, including sulfate, nitrate, and sea salt), is not provided by the TROPOMI L2 NO₂ product and is not included in the present analysis.
- **Radiative Transfer and Photochemistry:** 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. Solar zenith angle also affects the actinic flux, which is key in photochemistry: it directly drives NO₂ photolysis, and photochemistry in turn affects the levels of O₃ (which converts NO into the NO₂ that TROPOMI observes) and OH (a sink of NO₂). Total column water vapor affects NO₂ spectral signatures; ~~while top-of-atmosphere.~~ The TOA incident solar radiation ~~quantifies the energy available for the measurement~~ is the downward shortwave solar flux incident at the top of the atmosphere; it is strongly correlated with the solar zenith angle and primarily encodes the joint dependence on latitude and day-of-year.

2.4.4 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.~~ three albedo features describe surface reflectivity at different wavelengths. Two are at 758 nm, the wavelength of the oxygen A-band used by the FRESCO cloud retrieval: a surface albedo (taken from the TROPOMI Directional Lambertian Equivalent Reflectivity (DLER) climatology of Tilstra et al. (2024) and used as input to FRESCO) and a scene albedo (the effective Lambertian reflectivity of the scene treated as a single uniform reflector). The third is a surface albedo evaluated at 440 nm and applied across the NO₂ fitting window (405–465 nm) for both the cloud fraction retrieval at this wavelength and the air-mass factor calculation.

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.

270 Throughout this work we distinguish three related quantities that are easily conflated:

- **Detection:** a binary label produced by the Automated Plume Detection algorithm (Sect. 3.2) for a single TROPOMI observation: 1 if a statistically significant NO₂ enhancement is attributable to the target power plant in the downwind sector, 0 otherwise.
- **Detectability:** the probability of detection $P(\text{detect} | \mathbf{x})$ given a feature vector $\mathbf{x}$ that characterizes the observation conditions, abbreviated $P(\text{detect})$ when context is clear. Detectability is estimated either empirically (as the per-plant fraction of observations with a detection; Sect. 4.2) or via a machine learning model trained on the binary detection labels (Sect. 3.4).
- **Detection frequency:** the absolute per-plant count of detections over an observation period (e.g., number of detections per year).

280 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-Detection** 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:

285 **U.S. Analysis:** Uses hourly NO_x emission data (~~2019–2024~~2019–2024) from 500 high-emitting power plants with **Continuous Emissions Monitoring Systems (CEMS)**. ~~CEMS. For ablations using annual emissions, we use CAMPD’s annual NO_x totals downloaded from the same database.~~

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

3.1 Pairing TROPOMI Observations with Emissions

290 For each analysis, we pair every TROPOMI overpass over a selected power plant with the corresponding emission record to form the input observations used by all subsequent steps. The TROPOMI-derived input features for the model are extracted from the single pixel closest to the plant’s center, not aggregated across the plume.

U.S. The U.S. analysis uses 666,222 TROPOMI overpasses across our 500 selected plants. This full set of observations is used throughout the manuscript for dataset characterization and detection statistics. For the model training and feature-importance
295 analyses (Sections 3.4 and 3.5), we use 189,713 observations from 171 plants, namely the subset of overpasses that have (i) no missing (NaN) values in any input feature, since TROPOMI retrieval variables can themselves contain NaNs, (ii) a paired hourly NO_x record from the CAMPD database, and (iii) a target plant lying outside interference zones (Sect. 3.3). CAMPD

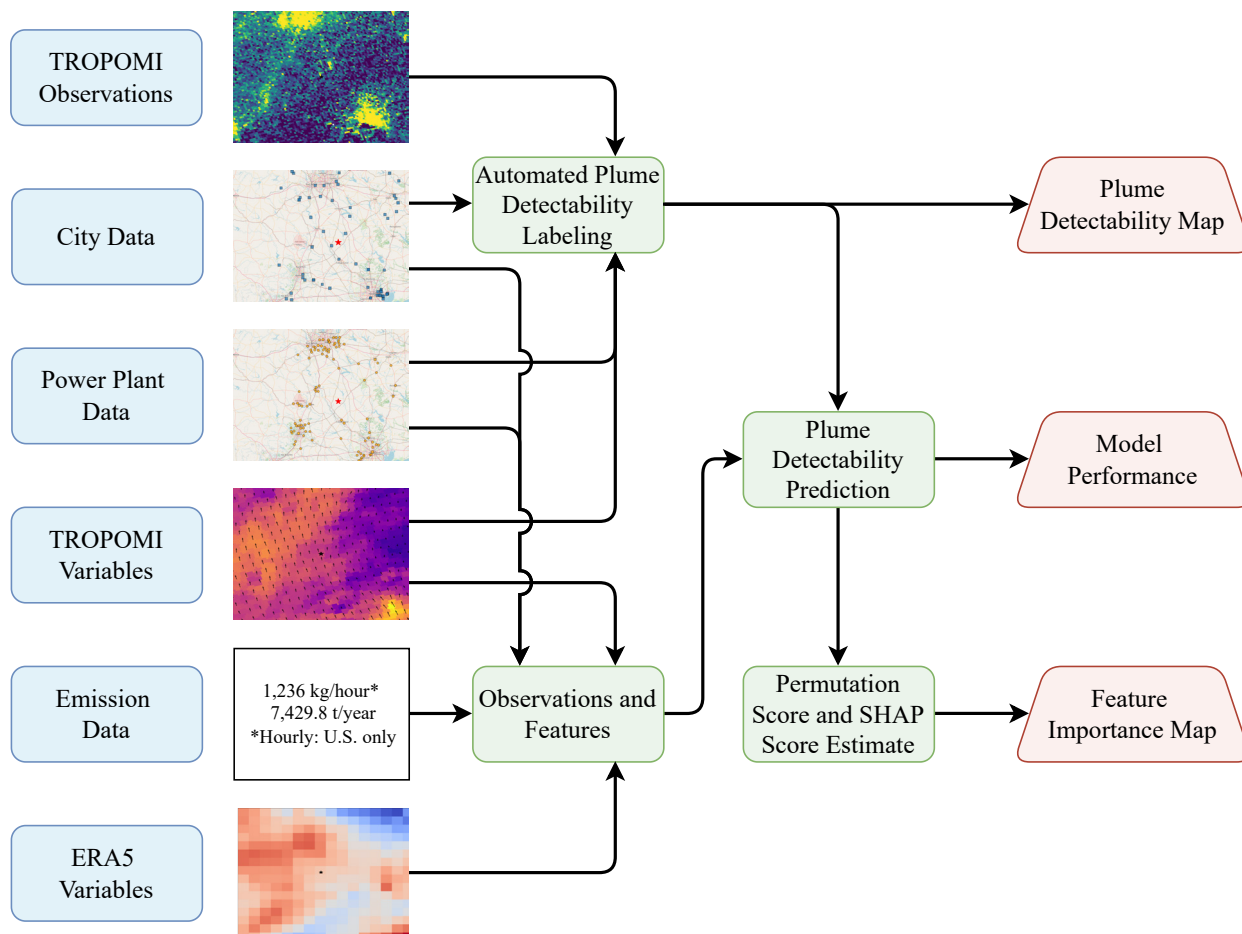


Figure 3. Overview of the analytical pipeline. The process uses input data (blue) to feed the main pipeline processes (orange). First, the Automated Plume Labeling-Detection algorithm produces detectability-binary detection labels for each observation. These labels, combined with extracted features, 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-Estimation). The final products (red) originate from these distinct stages: the labeling-detection 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.

only catalogues regulated facilities during operating hours with valid CEMS data, so overpasses during monitoring downtime, plant shutdowns, or over facilities below reporting thresholds are excluded.

300 **Global.** The global analysis uses 875,686 TROPOMI overpasses across our 6,000 selected plants. As with the U.S. dataset, this full set of observations is used throughout the manuscript for dataset characterization and detection statistics; for the model training and feature-importance analyses (Sections 3.4 and 3.5), we use 161,118 observations from 1,065 plants, namely the subset of overpasses with no missing (NaN) values in any input feature, a valid annual NO_x record in the CoCO₂ catalog, and a target plant lying outside interference zones (Sect. 3.3).

305 3.2 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~~ the Data-Driven Emission Quantification (DDEQ) toolkit (Kuhlmann et al., 2024) detect NO_2 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~~ plume-detection-with-attribution algorithm that automatically labels plumes and assigns each one to its originating power plant.

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.

315 3.2.1 Hyper-parameter Tuning

~~To tune the~~ The tuned hyper-parameters of the Automated Plume ~~Detectability Algorithm~~, Detection Algorithm are: in Step 1, the city population threshold, the plant emission-ratio threshold, the City Masking radius, and the Power Plant Masking radius; in Step 2, the wind-direction tolerance, the maximum search distance, the close-range distance, and the minimum plume area; in Step 4, the upwind sector angle and the background annulus; and in Step 5, the Statistical Significance (2σ) and Absolute Minimum thresholds. Their final values are reported in the corresponding step descriptions below. ~~To tune them,~~ we drew a ~~stratified sample of tuning set of 200 observations:~~ 100 ~~snapshots~~ from the global dataset ~~—balanced—~~, stratified across six continents and five emission-level quantiles ~~by selecting roughly 3–4 snapshots per stratum—and an additional (0–20%, 20–40%, ..., 80–100%) with roughly 3–4 observations per (continent, quantile) cell; and~~ 100 from the U.S. dataset, stratified across the same five emission quantiles with roughly 20 observations per bin. Stratified sampling here means dividing the dataset into strata (defined by continent and emission quantile ~~for the global set, by emission quantile alone for the U.S. set~~) and then sampling equally from each stratum to ensure representative coverage. ~~Hyper-parameters were then~~ Final values were selected using heuristic physical approximations and insights from this tuning set.

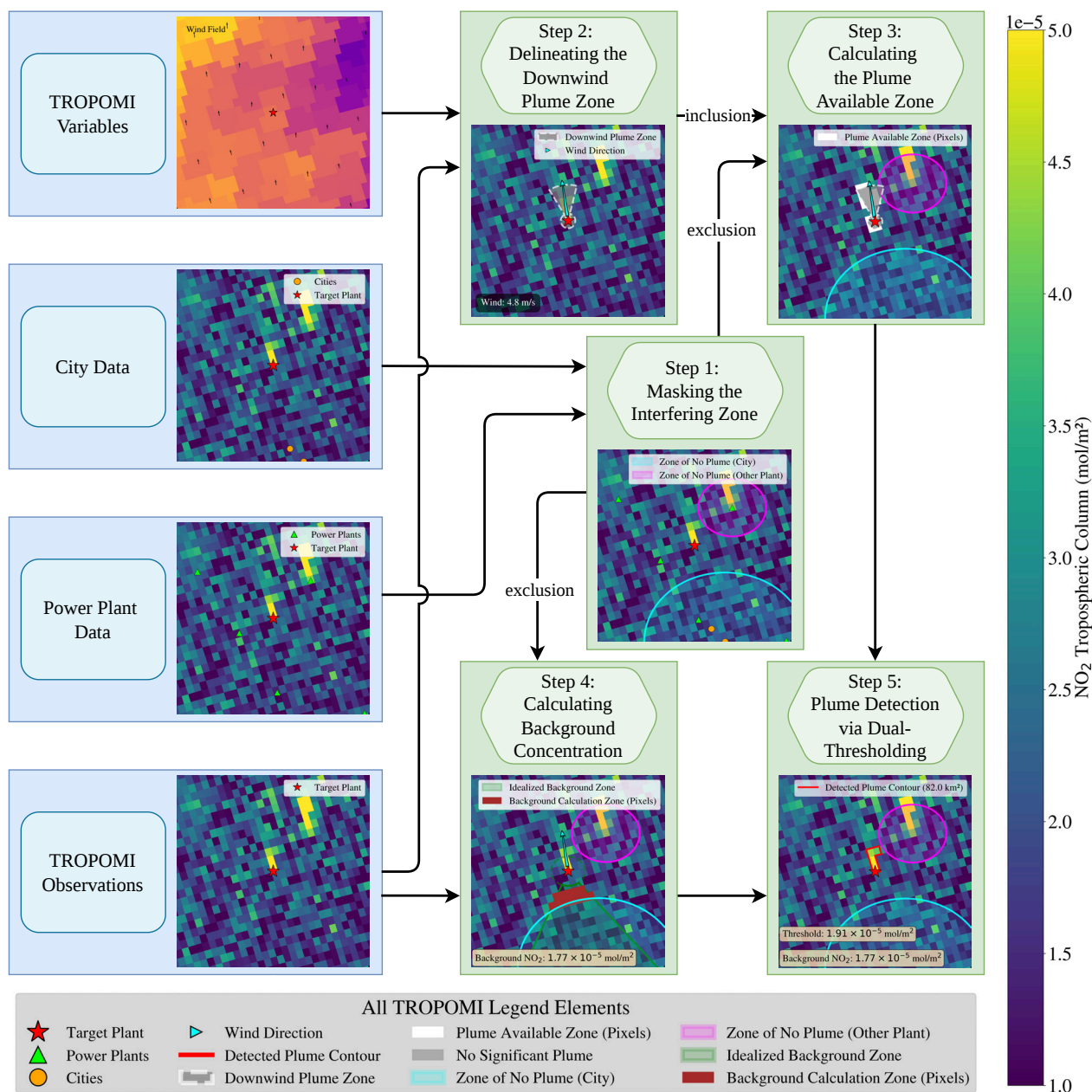


Figure 4. Overview of the Automated Plume Detection Algorithm. The process begins by masking potential interference from nearby cities and power plants (Steps 1-2). It then uses wind direction to define a downwind search area and isolates a “Plume Available Zone” by removing the masked regions (Steps 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.

3.2.2 Step 1: Masking Interference Zones

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~~ 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~~ 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 than the target plant's (an emission ratio of ~~1.0~~ 1.0 or higher); ~~the U.S. comparison uses each year's CAMPD-reported annual NO_x emissions, while the global comparison uses 2018 annual emissions from the CoCO₂ catalog.~~ 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 (Sect. 2.2) and the CoCO₂ catalog (Guevara et al., 2024) for global analysis ~~, mentioned in Section 3.~~ (Sect. 2.3).

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~~ 45 km and 90 km by applying a scaling factor of ~~9.0 from a zero~~ 9.0 from a zero baseline to the log population. This method ensures larger metropolitan areas are masked with a wider radius.
- **Power Plant Masking:** For other interfering power plants, the algorithm applies a fixed-radius mask of ~~20 km~~ 20 km. This standard distance excludes the immediate near-field zone where plumes from separate plants are most likely to mix and become indistinguishable.

The union of these city and power-plant masks defines the Zone of No Plumes, i.e., the area excluded from plume search in subsequent steps. This adaptive masking ensures that the subsequent analysis—, conducted within a ~~100-km window and characterized at 50 km—~~ 100 km window of the target plant, excludes pixels influenced by upwind pollution sources unrelated to the target plant. The mask is applied per scene to suppress pixels around interferers but does not preclude labeling the target plant itself: a plant whose centroid lies within an interference zone can still be detected in scenes where the wind direction places part of its downwind plume sector outside the masked region. Examples of city and power plant masking are shown in Appendix H.

3.2.3 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 100 m wind vectors (speed and direction) ~~directly from the TROPOMI satellite product for the from ERA5 reanalysis at the~~ time of the overpass.

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 .~~ Within 5 km of the plant, we relax the wind-direction tolerance to include any pixel containing the plant, since at this close range a pixel may contain the plant even when its centre lies outside the strict downwind cone. Finally, any detected enhancement must cover a minimum area of 25 km^2 to be flagged as a valid signal.

3.2.4 Step 3: Calculating the Plume Available Zone

The ~~Plume Available Zone~~ 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~~ Downwind Plume Zone from Step 2 and subtracting the ~~Zone of No Plumes~~ 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.

3.2.5 Step 4: Calculating Background Concentration

To accurately quantify a plume's enhancement, the algorithm first establishes the ambient tropospheric NO_2 ~~concentration~~ column density 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).

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~~ column density as the median of the remaining filtered pixels, and computes their standard deviation σ as a measure of local background variability for use in Step 5. Using the median ensures the background value is robust against outliers, providing a reliable baseline.

3.2.6 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 Zone. The algorithm flags a pixel as part of the plume only if its tropospheric NO_2 ~~concentration-meets-both~~ column density meets both of the following conditions:

1. **Statistical Significance:** The pixel's value must be at least ~~2 standard deviations (2σ)~~ 2 standard deviations (2σ) above the local background ~~concentration~~ column density 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 ~~$5 \times 10^{-6} \text{ mol/m}^2$~~ $5 \times 10^{-6} \text{ mol m}^{-2}$. This threshold prevents 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~~ 2.3 times 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) Van Geffen et al. (2020), to robustly distinguish real signals from sensor artifacts.

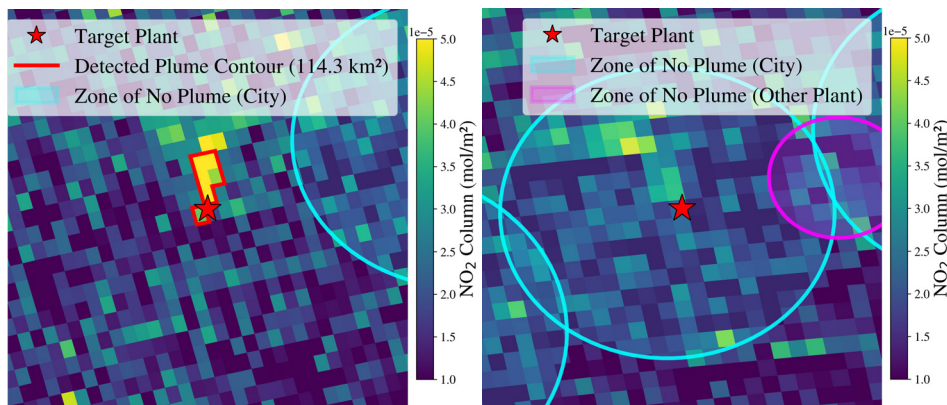
By requiring a signal to be both statistically significant relative to the background and above a fixed minimum, this dual-threshold method robustly identifies the emission plume while minimizing noise-based false detections. These binary labels indicate the presence of a NO₂ enhancement attributable to the plant; they do not require a coherent multi-pixel plume of the kind used for emission quantification.

3.2.7 Validation of the Plume **Labeling Detection** Algorithm via Manual Labeling

To validate the plume-labeling-plume-detection 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 hyperparameter tuning. The first test focuses on the U.S. dataset, where hourly emissions provide the cleanest reference; the second extends the assessment to the global dataset across continents.

In the first test, we evaluated algorithm performance under realistic, operational conditions using U.S. performance using an emission-stratified validation set of 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%) U.S. observations, with 80 observations per emission-level quantile defined above. Manual verification of this set provides a benchmark for real-world performance. The resulting confusion matrix (Figure 5c) shows U.S. performance across the emission range. Figure 5c shows the resulting confusion matrix, with an overall accuracy of 98.25%–98.0% (precision 93.5%, recall 93.5%).

In the second test, we assessed robustness by constructing a balanced-continent-and-emission-stratified global 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 observations. Observations were drawn from the global dataset by jointly stratifying across the same five emission-level quantiles (0–20%, 20–40%, etc.). This yielded a dataset with and across six continents (roughly 13 observations per (continent, quantile) cell, yielding 80 samples in each emission bin, ensuring no single region or emission category dominated observations per emission bin); we drop Antarctica because it contains no power plants in our inventory, and observations whose ISO3 country code does not map to one of the six continents are excluded prior to sampling. On this balanced-stratified global set, the algorithm achieved an accuracy of 97.75% (Figure 5d)–98.0% (precision 88.0%, recall 95.7%), as shown in Figure 5d.



(a) Positive detection

Validation on Random Sample ($n=400$). Results: 98.25%

Accuracy, 96.67% Precision, 82.86% Recall:

	Pred. False	Pred. True	Total
Actual False	364/334	1/4	365/338
Actual True	6/4	29/58	35/62
Total	370/338	30/62	400

(b) Negative detection

Validation on Stratified Sample ($n=400$). Results: 97.75%

Accuracy, 89.58% Precision, 91.49% Recall:

	Pred. False	Pred. True	Total
Actual False	348	5/6	353/354
Actual True	4/2	43/44	47/46
Total	352/350	48/50	400

(c) U.S. validation ($n=400$): 98.0% accuracy, 93.5% precision, 93.5% recall.

(d) Global validation ($n=400$): 98.0% accuracy, 88.0% precision, 95.7% recall.

Figure 5. Validation of the Automated Plume Detection Algorithm. (a) An A positive-detection example of a positive-detection, where: the algorithm correctly identifies a plume from a attributable to the target power plant after masking interference. (b) An A negative-detection example of a negative-detection: the target plant's location lies inside an interference-zone mask (no-plume zone) defined in Step 1, where so the algorithm returns no plume is attributed to-. Outside the target due-mask, a detection would additionally require the dual-threshold conditions of Step 5 (statistical significance and absolute minimum) and the minimum plume-area condition of Step 2 to interfering sources be met. (c, d) Confusion matrix-matrices for the U.S. and global validation on samples. "True" denotes the presence of a plume attributable to the target plant and "False" denotes its absence; "Predicted" is the algorithm's binary output and "Actual" is the manual human label assigned to the same observation. Green cells are correct classifications (TN and TP) and red cells are misclassifications (FP and FN). (c) The U.S. validation uses 400 randomly sampled records observations stratified across five emission quantiles (~ 80 per bin). (d) Confusion matrix for the The global validation on uses 400 records selected via observations jointly stratified sampling to ensure geographic across six continents and emission-level balance five emission quantiles (~ 13 per cell, 80 per emission bin). Dataset: two manually labeled validation sets ($n = 400$ U.S., $n = 400$ global) sampled from the full quality-filtered dataset ($n = 666,222$ U.S./ $n = 875,686$ global TROPOMI observations).

420 3.3 **Plume Detectability Prediction**

3.3 Training Sample Construction

In this section, we evaluate how effectively our input features distinguish between detectable and non-detectable plumes (plume detectability). Beyond the per-scene masking applied within the labeling algorithm (Sect. 3.2), we apply a plant-level filter to construct the dataset for the second-stage detectability analysis: we retain only plants that lie outside interference zones across every analyzed year, giving a cleaner dataset. To do so, we conduct an **item-split experiment**, training a model on a random sample of observations from all power plants.

We first 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 number of active plants 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 45.0% of emissions from the original 500.

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.

These contrasting retention rates, 45.0% 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 an observation level, requiring (i) no missing values in any input feature, (ii) a valid paired NO_x record (CAMPD for the U.S., CoCO₂ for global), and (iii) a target plant outside interference zones, the final modeling subset retains 189,713 of 666,222 U.S. observations (71.5% excluded across 2019–2024) and 161,118 of 875,686 global observations (81.6% excluded). To support the per-emission-magnitude analyses described later (Sect. 3.4), we also consider three subsets defined by the highest-emitting plants within each region: the Top 100, Top 50, and Top 20 by total NO_x emissions. Table 2 reports removal rates for these smaller U.S. subsets, which are similarly large.

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 NO_2 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.

450 3.4 **Plume Detectability Prediction**

Using the dataset constructed in Sect. 3.3, we train a machine learning model to predict plume detectability from the input features. We frame plume detectability as a binary classification task. For each TROPOMI observation, which our algorithm

Table 2. Subset-level observation filtering. “Retained” is the modeling subset, i.e., observations with no missing input features, target plant outside interference zones, and (for U.S. rows) a paired CAMPD record. Dataset: full quality-filtered dataset ($n = 666,222$ U.S. and $n = 875,686$ global TROPOMI observations), with each “Top X ” row restricted to the Top X NO_x emitters.

Subset	Unfiltered	Retained	Removed	Removal Rate
All (Global)	875,686	161,118	714,568	81.6%
Top 100 (Global)	17,460	6,110	11,350	65.0%
Top 50 (Global)	8,672	3,206	5,466	63.0%
Top 20 (Global)	3,303	797	2,506	75.9%
All (U.S.)	666,222	189,713	476,509	71.5%
Top 100 (U.S.)	133,414	65,930	67,484	50.6%
Top 50 (U.S.)	66,939	40,667	26,272	39.2%
Top 20 (U.S.)	27,028	17,820	9,208	34.1%

labels as either ~~“plume detected”~~ or ~~“plume detected”~~ or “no plume,” we extract a comprehensive set of meteorological, ~~instrumental~~ sensor, environmental, and ~~emission-related~~ power-plant features to use as classifier inputs.

455 We ~~selected~~ use a Multi-Layer Perceptron (MLP) ~~after finding it outperformed random forests and logistic regression~~ for this binary classification task. 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~~ 25 input features (24 numeric features plus the integer-encoded primary fuel type) and produces a binary classification output. ~~With a 25-dimensional input, the model has 53,761,~~ giving
460 ~~49,921~~ 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”~~ “All” (our complete power plant inventory) and ~~three subsets of the highest emitters (“Top-100”, “Top-50”, and “Top-20”).~~ the three highest-emitter subsets (“Top 100”, “Top 50”, and “Top 20”) defined in Sect. 3.3. 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 3.2). 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 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
470 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 C.

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 for the We evaluate two complementary splits: an item split, in which observations are randomly assigned to train, validation, and test sets (in-distribution generalization), and a power plant split experiment are presented in Appendix F (Table F1). 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: in which entire plants are held out for testing (out-of-distribution generalization to unseen facilities). For each model we report six standard classification metrics (Accuracy, 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). AUC, and Cohen’s κ); item-split results are presented in Sect. 4.3 and power-plant-split results in Appendix F.

3.5 Permutation Score and SHAP Score ~~Estimate~~ Estimation

3.5.1 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 the held-out validation set $\mathcal{D}_{\text{val}}$ test set $\mathcal{D}_{\text{test}}$, 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 examples in $\mathcal{D}_{\text{val}}$ $\mathcal{D}_{\text{test}}$ 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.

3.5.2 Plant-Level Feature Importance via SHAP

While permutation importance gives a global view, we use SHAP (SHapley Additive exPlanations) to explain predictions for individual power plants in the held-out test set $\mathcal{D}_{\text{test}}$. 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 $\mathbf{x} \in \mathbb{R}^d$ ($d = 25$ in our case), 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); where $\mathbf{X}$ is the random feature vector over the dataset and $\mathbf{x}$ is a specific plant’s realization:

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

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~~ tractable 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~~ detectability, 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.

To visualize these local explanations on a global scale, we aggregate the plant-level SHAP results onto a grid, as shown in Figure D2. 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

4.1 ~~Most power plants are near other emission sources~~ Scope of the analysis

~~First, we find that for the~~ After interference filtering (Sect. 3.3), the analyzed set consists of 171 U.S. top-500 NO_x plants outside interference zones (retaining 45.0% of the original 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.

~~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 Top 500 selection) and 1,065 plants remained global plants outside interference zones, retaining just (retaining 21.1% of the original emissions.~~

~~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 NO_x emissions from the Top 6,462 of 875000 selection); 189,580 TROPOMI snapshots (81.8%), retaining 161713 of 666,118. For the U.S. dataset, 305,827 of 504,165 snapshots (60.7%) were removed across 2019–2024. Cohort-level aggregates for smaller 222 U.S. subsets show similarly large removals (Table ??); per-year details for the top-500 are given in Table A1.~~

535 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.

540 Cohort-level snapshot filtering (aggregated, 2019–2024): Cohort Unfiltered Retained Removed Removal Rate All (Global) 875,580 observations and 161,118 of 875,462 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% 686 global observations remain. All detectability results in the following subsections refer to this filtered set.

4.2 NO₂ plume detectability has regional disparities

4.3 NO_x plume detection probabilities have regional disparities

With power plants near interfering emitters removed, we proceeded with a clean set of snapshots observations 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 11 to 258 per plant on average each year, with a median of 46.5 48. This range is highly uneven: plants in the 90th percentile (152.33 at P₉₀ (154) record about 8.38× 8.6× more detections per year than those in the 10th percentile (18.17 at P₁₀ (18)). 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) the per-plant empirical detectability spans from 6.0% to 85.2% (median 20.4%; P₁₀ = 10.0%, P₉₀ = 63.7%).

Globally (1,065 plants after filtering), detections over the eight-month period in 2018 range from 0 to 248 251, with a median of 37.0. Plants in the 90th percentile (140.738. Plants at P₉₀ (141) record about 11.73× 11.8× more detections than those in the 10th percentile (12.0 at P₁₀ (12)). When adjusted for the number of satellite overpasses, detection probabilities span the per-plant empirical detectability spans from 0% to 100 per 100 observations (median 27.48 per 100; P₁₀=10.25, P₉₀=80.73% (median 27.8%; P₁₀ = 10.5%, P₉₀ = 81.8%).

These disparities have a clear geographical distribution: detection probabilities are detectability is 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 also strongly affect plume detectability ~~as well (Fig. 6a,c), as we will see in the next section,~~ as visible in panels (a) and (c) of Figure 6; we revisit these geographic patterns in Sect. 4.4 in light of the feature-importance results.

Despite this aggressive filtering, satellite detection ~~rates-frequencies~~ remain remarkably high ~~(Figure 6), as shown in Figure 6.~~
570 All 171 U.S. plants exhibit detection ~~rates-frequencies~~ exceeding 10 ~~snapshots-observations~~ per year, with 87.7% detected more than 20 times annually and nearly half (~~45.0~~46.2%) detected more than 50 times per year. The global dataset shows similarly strong performance, with ~~93.0~~93.3% of plants detected more than 10 times and three-quarters (~~74.5~~75.1%) detected more than 20 times during the eight-month observation period. ~~Notably, 23.4~~Figure 7 further shows that 24.6% of U.S. plants and ~~19.4~~19.8% of global plants exceed 100 positive detections in their respective observation periods, demonstrating TROPOMI's
575 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~~ Plume detectability is predictable with high accuracy across U.S. and global datasets

580 We report results from the two split designs introduced in Sect. 3.4. Table 3 reports the item-split results for both U.S. and global datasets; the power-plant-split results are reported in Table F1 of Appendix F. The tables organize performance by region, emission data resolution (annual vs. hourly NO_x), and dataset scope (All vs. Top *X* emitters). Of the six classification metrics reported, we focus discussion on Precision, Recall, and the F1 score, which together quantify the reliability of positive predictions, the ability to capture all true events, and the balance between these competing objectives in operational monitoring.

585 Our model predicts plume detectability from meteorological and emissions features with high accuracy ~~(. As reported in Table 3).~~ Across, across U.S. and global datasets, the multi-layer perceptron reaches ~~0.740, 75–0.76~~0.78 accuracy, 0.65–0.67 F1 score and achieves AUCs above 0.80; the underlying ROC curves for the “All” subset are shown in Figure D1 of Appendix D1. These metrics show that sensor, power plant, meteorological, and environmental variables can predict plume detectability,
590 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, as seen by comparing the “All”, “Top 100”, “Top 50”, and “Top 20” rows of Table 3, where each row corresponds to a separate model trained and evaluated on the indicated subset (Sect. 3.4). As the subset is restricted to higher emitters,~~ Precision, Recall,
595 and the F1 score all rise, with Precision showing the ~~most significant largest~~ improvement. This ~~improvement indicates that trend suggests that on these subsets~~ the model makes fewer false positive errors ~~when identifying these stronger-, since stronger emitters tend to produce~~ 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

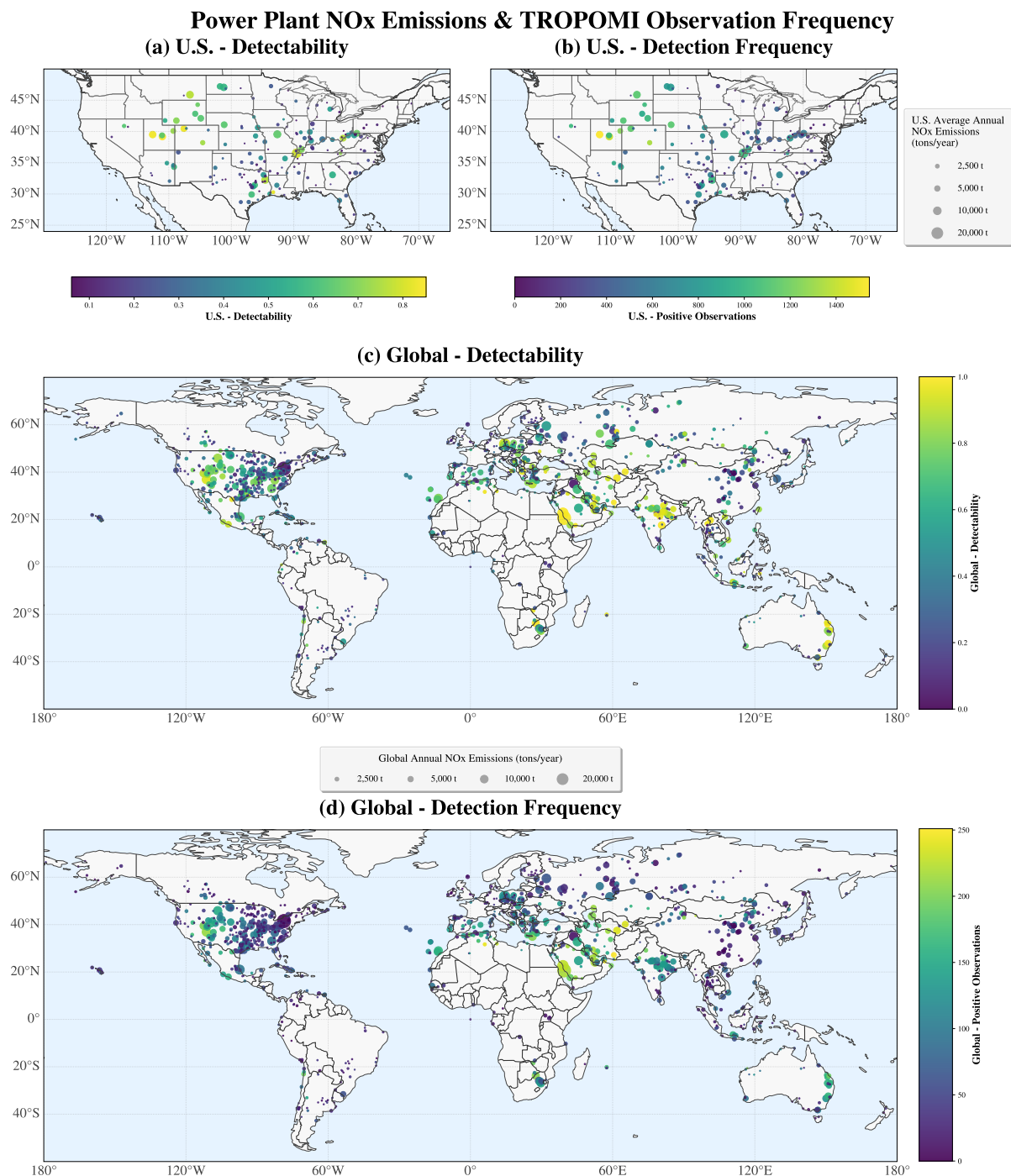


Figure 6. Maps of NO_x emissions and TROPOMI observations: marker size $\propto$ annual NO_x (metric t/year yr^{-1}) at each plant, with U.S. panels colored by (a) plume ~~detection-probability~~detectability and (b) ~~positive-observation-count~~detection frequency, and global panels colored by (c) plume ~~probability~~detectability and (d) ~~positive-observation-count~~detection frequency. Dataset: per-plant aggregate over the full quality-filtered dataset, restricted to interference-free plants with no missing-feature filter applied (171 U.S. plants, $n = 234,038$ observations; 1,065 global plants, $n = 161,143$ observations).

TROPOMI NO_x Plume Detection Frequency vs Annual Emissions - Grid Analysis

(a) United States Power Plants
2019-2024

(b) Global Power Plants
2018

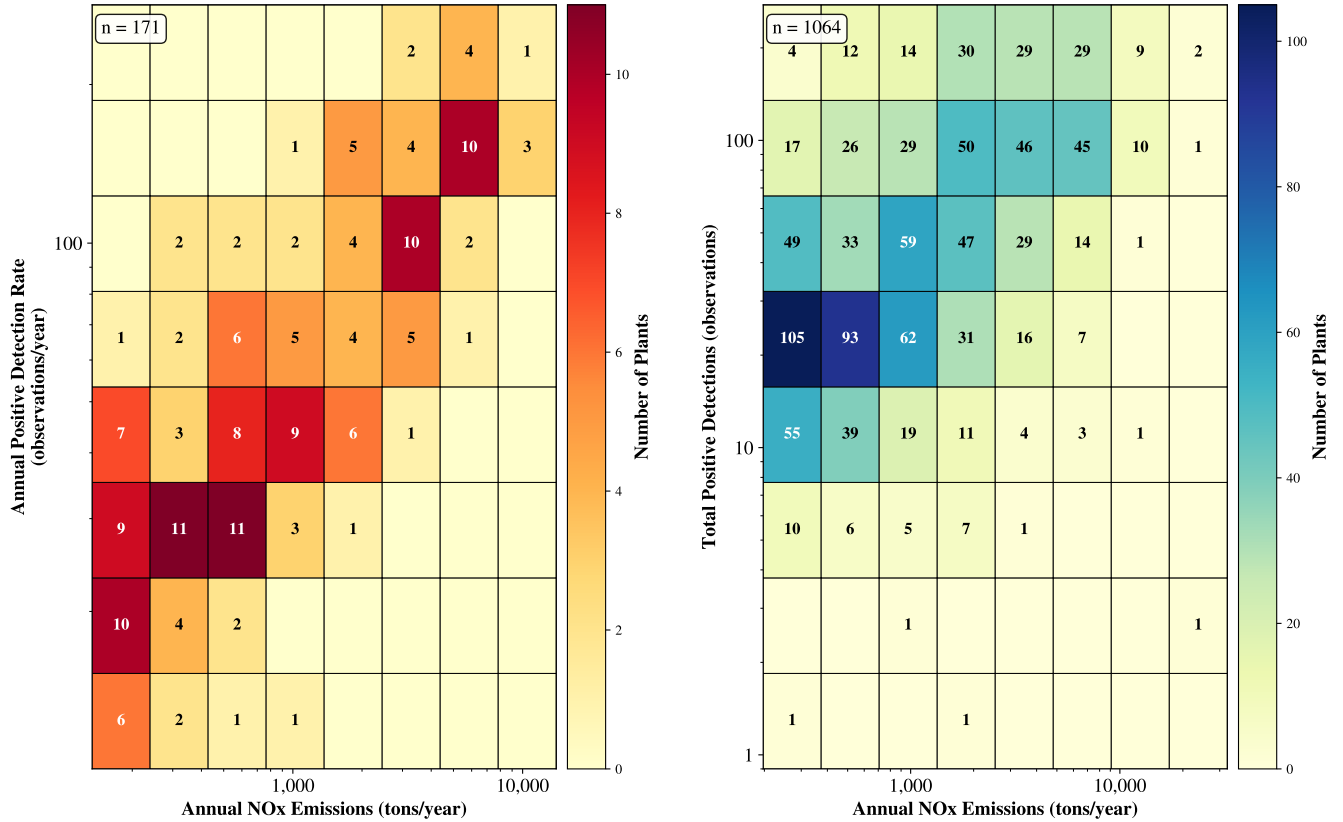


Figure 7. TROPOMI NO_x detection frequency versus annual emissions ~~after interference-zone filtering~~. (a) U.S. plants ($n = 171$, retaining ~~49.1~~45.0% of emissions from the original 500 plants). (b) Global plants (~~$n = 1065$~~ $n = 1,065$, retaining 21.1% of emissions from the original ~~6000~~6,000 plants). ~~The total number of~~ 1,064 plants ~~are plotted~~ is 1,064, as one plant yielded no positive detections. Numbers denote plant counts per grid cell (8×8 log-uniform bins). Dataset: per-plant aggregate over the full quality-filtered dataset, restricted to interference-free plants with no missing-feature filter applied (171 U.S. plants, $n = 234,038$ observations; 1,065 global plants, $n = 161,143$ observations).

identifies them, boosting the overall accuracy score. When the dataset is filtered to only the top emitters, these easier-to-classify negative cases are removed. ➔

Though the global dataset uses coarser temporal emission data, the model trained on it ~~performs similarly as~~ matches or exceeds the model trained on the U.S. dataset ~~across all subsets, with the gap widening for the highest emitters (e.g., on Top 20, Global F1 = 0.935 vs. U.S. hourly F1 = 0.825)~~. 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~~ helping the global-trained model generalize better.

Table 3. Multi-Layer Perceptron ~~Performance~~ performance (item split) for U.S. and ~~Global~~ global datasets. Dataset: modeling subset (Sect. 3.3; 171 U.S. plants, $n = 189,713$ observations; 1,065 global plants, $n = 161,118$ observations); each “Top X ” row uses the corresponding subset of Table 2.

Region	Emissions	Dataset	Accuracy	Precision	Recall	F1 Score	AUC
U.S.	Annual NO _x	All	0.750 ± 0.004 <u>0.757 ± 0.003</u>	0.619 ± 0.009 <u>0.638 ± 0.006</u>	0.669 ± 0.010 <u>0.653 ± 0.007</u>	0.643 ± 0.002 <u>0.645 ± 0.004</u>	0.799 ± 0.003 <u>0.803 ± 0.002</u>
		Top-100 <u>Top 100</u>	0.707 ± 0.003 <u>0.717 ± 0.003</u>	0.757 ± 0.012 <u>0.772 ± 0.006</u>	0.713 ± 0.020 <u>0.714 ± 0.006</u>	0.734 ± 0.005 <u>0.742 ± 0.003</u>	0.777 ± 0.002 <u>0.787 ± 0.001</u>
		Top-50 <u>Top 50</u>	0.702 ± 0.007 <u>0.710 ± 0.004</u>	0.801 ± 0.019 <u>0.818 ± 0.002</u>	0.703 ± 0.020 <u>0.708 ± 0.007</u>	0.749 ± 0.009 <u>0.759 ± 0.004</u>	0.771 ± 0.006 <u>0.779 ± 0.003</u>
		Top-20 <u>Top 20</u>	0.710 ± 0.003 <u>0.720 ± 0.012</u>	0.856 ± 0.005 <u>0.879 ± 0.008</u>	0.718 ± 0.004 <u>0.719 ± 0.023</u>	0.781 ± 0.003 <u>0.791 ± 0.012</u>	0.771 ± 0.009 <u>0.792 ± 0.004</u>
	Hourly NO _x	All	0.763 ± 0.004 <u>0.775 ± 0.003</u>	0.645 ± 0.012 <u>0.670 ± 0.009</u>	0.660 ± 0.018 <u>0.666 ± 0.015</u>	0.653 ± 0.003 <u>0.668 ± 0.005</u>	0.809 ± 0.002 <u>0.819 ± 0.001</u>
		Top-100 <u>Top 100</u>	0.733 ± 0.004 <u>0.751 ± 0.002</u>	0.781 ± 0.008 <u>0.807 ± 0.004</u>	0.736 ± 0.008 <u>0.740 ± 0.009</u>	0.758 ± 0.004 <u>0.772 ± 0.004</u>	0.804 ± 0.003 <u>0.820 ± 0.001</u>
		Top-50 <u>Top 50</u>	0.734 ± 0.003 <u>0.754 ± 0.005</u>	0.816 ± 0.006 <u>0.854 ± 0.003</u>	0.747 ± 0.015 <u>0.746 ± 0.011</u>	0.780 ± 0.006 <u>0.796 ± 0.006</u>	0.805 ± 0.003 <u>0.833 ± 0.001</u>
		Top-20 <u>Top 20</u>	0.739 ± 0.009 <u>0.764 ± 0.008</u>	0.870 ± 0.006 <u>0.905 ± 0.007</u>	0.749 ± 0.019 <u>0.758 ± 0.013</u>	0.805 ± 0.010 <u>0.825 ± 0.007</u>	0.805 ± 0.004 <u>0.845 ± 0.001</u>
Global	Annual NO _x	All	0.745 ± 0.005 <u>0.746 ± 0.003</u>	0.679 ± 0.019 <u>0.684 ± 0.005</u>	0.652 ± 0.007 <u>0.648 ± 0.005</u>	0.665 ± 0.003 <u>0.665 ± 0.002</u>	0.800 ± 0.003 <u>0.801 ± 0.001</u>
		Top-100 <u>Top 100</u>	0.755 ± 0.006 <u>0.774 ± 0.010</u>	0.890 ± 0.007 <u>0.892 ± 0.007</u>	0.764 ± 0.013 <u>0.793 ± 0.024</u>	0.822 ± 0.005 <u>0.840 ± 0.011</u>	0.828 ± 0.009 <u>0.836 ± 0.001</u>
		Top-50 <u>Top 50</u>	0.793 ± 0.020 <u>0.805 ± 0.019</u>	0.893 ± 0.016 <u>0.897 ± 0.020</u>	0.814 ± 0.018 <u>0.831 ± 0.014</u>	0.852 ± 0.016 <u>0.863 ± 0.015</u>	0.864 ± 0.020 <u>0.877 ± 0.001</u>
		Top-20 <u>Top 20</u>	0.850 ± 0.031 <u>0.893 ± 0.027</u>	0.963 ± 0.013 <u>0.966 ± 0.014</u>	0.859 ± 0.040 <u>0.907 ± 0.025</u>	0.907 ± 0.020 <u>0.935 ± 0.016</u>	0.902 ± 0.036 <u>0.920 ± 0.001</u>

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~~. Figure 8 shows the geographic distribution and 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 industrialized regions including North America, Europe, and the Arabian Peninsula, with F1 scores ~~frequently exceeding~~ often reaching 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. The sparser training data in these regions is itself partly a consequence of factors that make satellite retrievals more difficult there, including high aerosol load, water vapor, and persistent cloud cover, which reduce the number of high-quality TROPOMI observations available for analysis. 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 (~~this relationship is~~ verified in Figure ~~D3~~) D3.

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 CO_2 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 t in the EPA dataset but 29,652 t in CO_2 —an 800-fold, an ~ 830 -fold discrepancy. If the EPA data is correct, the facility’s small actual size explains the poor detectability. In other words, the CO_2 dataset is likely incorrect about Domtar Paper Company’s emissions.

4.4 Which features are most predictive of plume detectability?

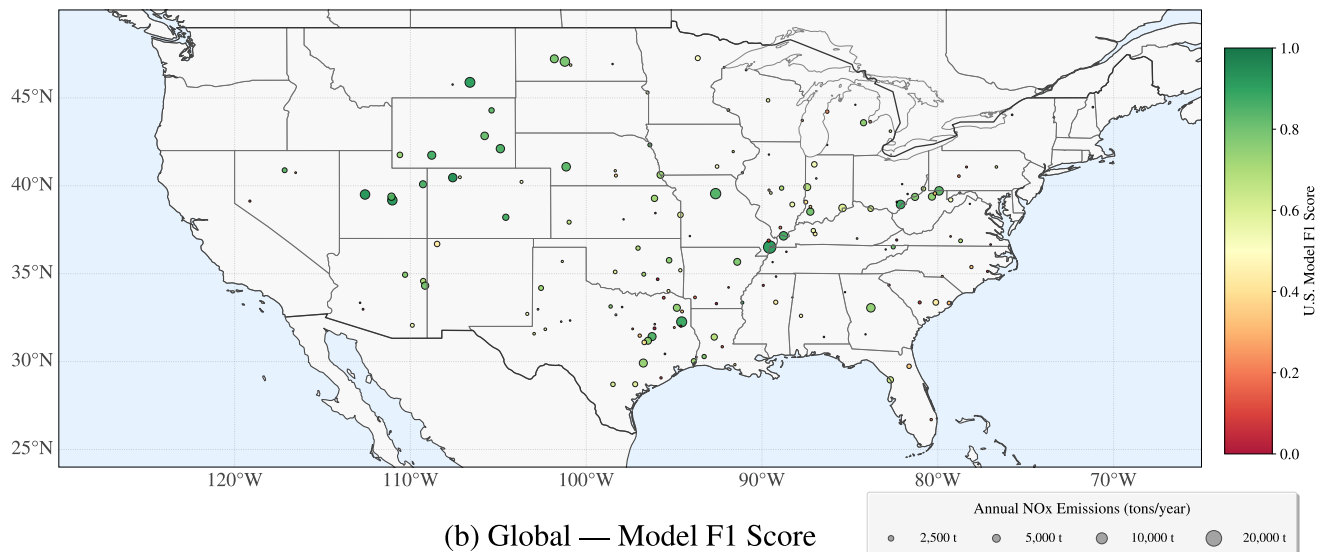
To identify the key predictors of NO_2 plume detectability, we performed a permutation importance analysis on our best-performing models (Table 3). Figure 9 (left) shows the ranked feature importance, sorted by the combined U.S. and global importance. The top 5 environmental and power plant-specific features include features are NO_x emission rate, surface albedo, surface albedo (NO_2 window), sensor zenith angle, and primary fuel type, and wind speed (Figure 9, left), with wind speed close behind at rank 6. Our SHAP score analysis yielded similar results, which are shown in Appendix D2. A complete list of permutation-importance values appears in Appendix G, Table G1. Cloud fraction appears low in this ranking, but still shapes detectability within the post-ga range; we show this in Appendix G2.

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 20 quantile bins for each feature ($\approx 7,800$ samples/observations/bin globally; $\approx 9,700$ – 300 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), are shown in the right panel of Figure 9.

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). This albedo dependence is consistent with the geographic patterns reported in Sect. 4.2: arid and semi-arid regions have higher detectability partly because of their elevated surface albedo, while humid regions are penalized by both lower albedo and more frequent cloud cover. Our analysis finds that for U.S. power plants, an average hourly emission rate of $\approx 400 \text{ kg h}^{-1}$ corresponds to a $\sim 50\%$ detection probability. The result for global power plants is similar, although the global $P(\text{detection}) - P(\text{detect})$ 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^{-1} for the U.S., $\text{tons/year t yr}^{-1}$ for global). We align the x-axis for U.S. emissions (kg h^{-1}) and global emissions ($\text{tons/year t yr}^{-1}$) 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.

Per-Plant Model Performance: F1 Scores from MLP Models (paper revision)

(a) U.S. — Model F1 Score



(b) Global — Model F1 Score

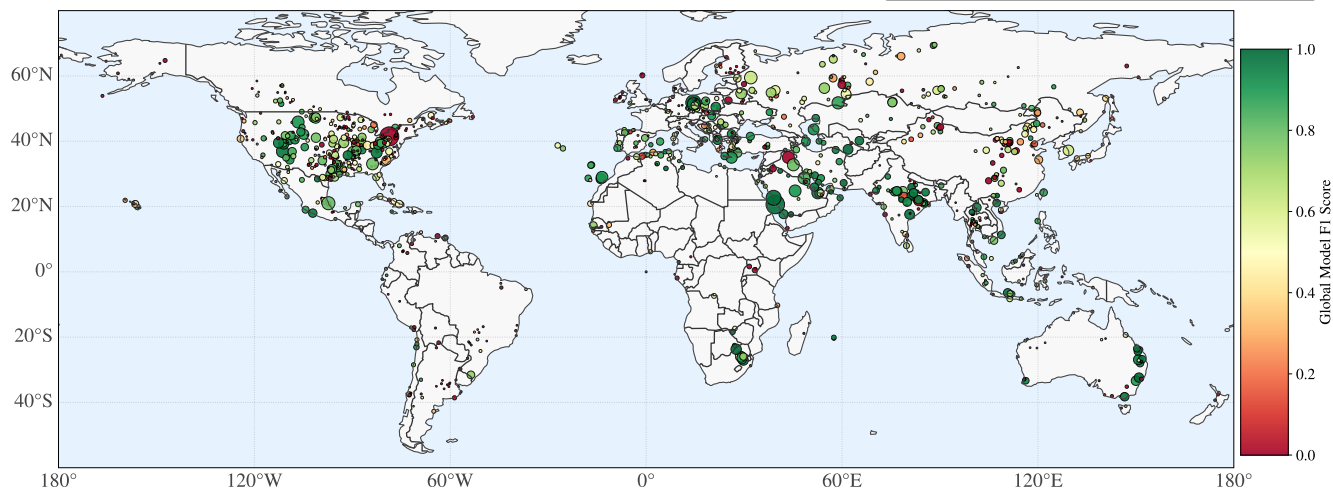


Figure 8. Geographic distribution of multi-layer perceptron (MLP) model performance in detecting NO_x plumes from individual power plants, evaluated using the F1 score. Panel (a) shows results for the U.S. model, while panel (b) presents the global model. Each point represents a power plant, with color 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-Midwest areas. The global model demonstrates strong performance across North America, Europe, and the Arabian Peninsula. For each plant, the F1 score is computed over all of its observations across the train, validation, and test splits. Dataset: modeling subset (Sect. 3.3; 171 U.S. plants, $n = 189,713$ observations; 1,065 global plants, $n = 161,118$ observations).



Figure 9. Feature importance and probability-of-detection-detectability analysis for TROPOMI-based TROPOMI-based NO₂ models at U.S. (red) and global (teal) scales. Left: ranked feature-permutation importance for the top 12 features, sorted by the sum of U.S. and global importance and computed on the held-out test split of the dataset. Right: detectability as a function of the top three features NO_x emission rate, surface albedo (NO₂ window), sensor zenith angle, and wind speed, computed on the dataset. Binning uses a quantile-based strategy (1st–99th 1st–99th percentile) to ensure balanced feature distribution and avoid extreme values. Dataset: modeling subset (Sect. 3.3; 171 U.S. plants, $n = 189,713$ observations; 1,065 global plants, $n = 161,118$ observations); the held-out test split contains $n = 37,943$ U.S. and $n = 32,224$ global observations.

By contrast, the sensor zenith angle (SZAVZA) shows a hump-shaped dependence. ~~From nadir to moderate off-nadir,~~
655 ~~detectability rises because longer slant paths at moderate SZA increase the tropospheric air-mass factor and the~~, with mean
peak plume detectability near 20° VZA globally and 35° in the U.S. Two effects can possibly explain the rising flank. First,
retrievals measure NO₂ slant column density along the line of sight ~~intersects more of the plume~~ and convert it to vertical
column density by dividing by the AMF; at higher VZA the longer slant path yields a larger AMF (Palmer et al., 2001), so the
same slant-column noise translates into a smaller vertical-column noise. Lower background noise raises detection probability
660 at fixed plume signal. Second, our 25 km² flagged-area threshold introduces a pixel-quantization step: TROPOMI pixel area
grows from ~20 km² at nadir to ~25 km² near VZA = 26° (Fig. G4), so detection requires > 2 contiguous flagged pixels
below 26° but only one above, producing a step increase in detectability. At larger VZA, ~~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 the larger pixel size dilutes the plume signal more strongly, and~~
665 ~~retrieval uncertainty increases~~ 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~~ detectability. We note that this is different from the ideal wind speeds for quanti-
fying emissions, which prior literature shows to occur at moderate wind speeds (2–4 m/s) s⁻¹ (Bruno et al., 2024). Very low
670 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). These two relationships are not in conflict because detection and quantification
address different tasks: quantification requires a coherent plume structure that the retrieval can fit, which breaks down at very
low wind (no advection); detection only asks whether any NO₂ enhancement is present at the source, which is favored by low
675 wind because NO₂ accumulates near the source.

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: ~~0.05~~0.049). This difference likely stems from the relative homogeneity of the U.S.
power sector in both its fuel sources and emission controls. First, the Top 500 U.S. ~~cohort of top-emitting power plants is~~
emitters are dominated by natural gas (~61%) and coal (~32%), which together account for ~93% of the facilities. This low
680 fuel variability is in stark contrast to the global dataset. While also led by coal (~50%) and natural gas (~30%), the global
mix is far more diverse, including substantial shares of facilities running on oil (~12%), biomass (~6%), and waste (~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
685 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 in Figure 10 five representative power plants that span the range of

conditions observed in our dataset (Figure 10). ~~Richmond County Plant, Marion~~, with low NO_x emissions ($\sim 72 \text{ kg/h}$ ~~83 kg h^{-1}~~) and low albedo (0.04), has the lowest ~~detection probability ($P(\text{detect}) = 0.13$)~~ detectability ($P(\text{detect}) = 0.18$). Ottumwa shows ~~moderate~~ higher NO_x ($\sim 185 \text{ kg/h}$ ~~180 kg h^{-1}~~) but experiences high wind speeds (4.24 m/s ~~and 5.74 m s^{-1}~~) and a low but highly variable albedo (0.07 ± 0.08), ~~reflected in its long-tail distribution mean ≈ 0.07 , with a long tail~~ from winter snow cover, ~~yielding $P(\text{detect}) = 0.24$~~ , yielding $P(\text{detect}) = 0.25$.

Comparing ~~Richmond County Marion~~ and Independence directly demonstrates the dominant effect of emission rate: Independence's high NO_x ($\sim 662 \text{ kg/h}$ ~~achieves $P(\text{detect}) = 0.56$~~ 670 kg h^{-1} achieves $P(\text{detect}) = 0.57$) despite similarly low albedo (0.05). Among the three high-emitting plants (Independence, Dave Johnston, ~~Bonanza~~ Huntington), surface characteristics and wind become the limiting factors. Independence's low albedo constrains detectability to ~~0.56~~ 0.57, while Dave Johnston (high wind, 5.01 m/s ~~reaches 0.63~~, ~~and Bonanza 6.69 m s^{-1}~~ ; albedo 0.09) ~~reaches 0.61~~, and Huntington (low wind, 2.74 m/s ~~3.81 m s^{-1}~~ ; high albedo, ~~0.13~~ 0.12) achieves the highest ~~detection probability at 0.73~~ detectability at 0.64. These examples demonstrate that high surface albedo and low wind enhance NO_2 retrieval sensitivity even when emission rates are comparable.

5 Discussion

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~~ estimate 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~~ plume-detection and machine-learning framework that links satellite detectability to meteorological, environmental, ~~and sensor-geometry~~ sensor, and power-plant features. Our goal 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 angles enhance detectability by increasing the optical path length through the lower troposphere, consistent with established ~~air-mass-factor~~ AMF 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).

Our analysis extends these theoretical principles by quantifying their global impact empirically. Across thousands of facilities, detectability peaks at intermediate sensor zenith angles ~~of 25° to 30° and~~ (near 20° globally and 35° in the U.S.) 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 $\approx 50\%$ ~~probability of detection~~detectability. These results provide the first global observational confirmation of how viewing geometry, surface brightness, meteorology, and emission strength jointly control satellite plume visibility.

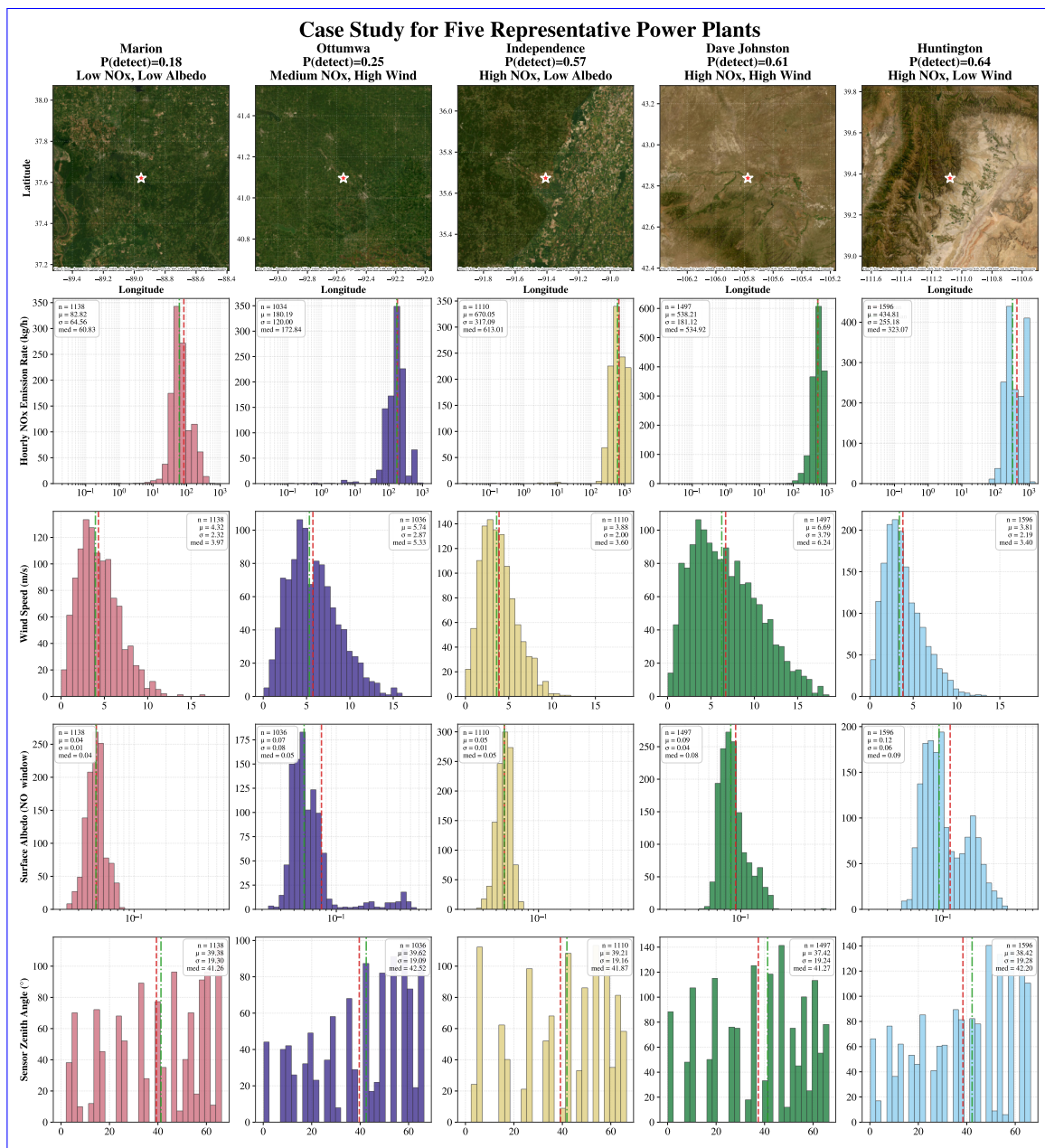


Figure 10. Variable distributions for five representative power plants demonstrating detectability drivers. Plants were selected ~~from 171 consistently monitored power plants in the U.S.~~ to showcase diverse NO_x emissions, meteorological conditions, and viewing geometries. Top: Satellite imagery with ~~detection probabilities~~ detectability values and dominant characteristics. Rows 2-5: Histograms of hourly NO_x emission rates (kg/h^{-1}), wind speed (m/s^{-1}), 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 Basemap: Esri, i-cubed, USDA, USGS, AEX, GeoEye, Getmapping, Aerogrid, IGN, IGP, UPR-EGP, and ArcGIS World Imagery (Esri, n.d.). Dataset: five U.S. plants drawn from the GIS User Community | Powered by Esri 171 plants of the U.S. modeling subset (Sect. 3.3; $n = 189,713$ observations); each histogram uses the selected plant's observations from this subset.

725 5.2 Implications for emission retrieval

The same meteorological, environmental, ~~and geometric-sensor, and power-plant~~ 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 machine-learning-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.

We emphasize that detectability as defined here is necessary but not sufficient for emission quantification: it characterizes the upstream observational limit (whether any NO₂ enhancement attributable to the plant is present in the scene), whereas quantification additionally requires that the plume’s structure be sufficiently coherent to recover an emission rate. The conditions favorable for detection and for quantification therefore overlap but are not identical; for example, low wind favors detection (because NO₂ accumulates near the source) but is unfavorable for quantification methods that require a clearly advected plume.

5.3 Robustness to inventory uncertainty

A natural concern is that the reported emission rates used as a predictor are themselves uncertain. Reported inventories disagree both with satellite-derived estimates (e.g., Beirle et al. 2019; Goldberg et al. 2019) and among themselves (e.g., EDGAR (Crippa et al., 2020) vs. E-PRTR (European Environment Agency, 2020)). Within the global subset we use CoCO₂ annual emissions (Guevara et al., 2024), while the U.S. subset uses hourly emissions (U.S. Environmental Protection Agency, 2023).

Two factors degrade the signal the model can extract from the reported emissions: noise in the reported inventories, and the use of annual emissions for the global subset (the only data available at global scale), which is a less faithful proxy of the emissions at each observation. The reported F1 and AUC values should therefore be read as lower bounds on the model performance achievable under cleaner inputs. The same noise feeds into our feature-importance estimate, and the corresponding feature appears less important than it truly is. The top SHAP ranking of NO_x emission rate is therefore a conservative lower bound, and access to cleaner inventories, including global hourly data, would only reinforce its dominance.

In this work, we systematically mapped power plant NO₂ plume detectability at U.S. and global scales using TROPOMI observations (nadir pixel size 3.5–7 km), and then demonstrated that detectability can be predicted by a suite of ~~environmental and sensor variables~~ meteorological, environmental, sensor, and power-plant variables, with our trained models achieving F1 score > 0.66 and AUC > 0.8 across U.S. and global datasets. 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~~ rate, surface altitude, surface albedo (NO₂ window), sensor zenith angle, primary fuel type, and wind speed. Because the model is trained on TROPOMI features, its predictions apply to TROPOMI-like observations rather than to satellite NO₂ retrievals in general. Because the analysis is restricted to plants outside interference zones (at least 20 km from other major power plants and 45–90 km from cities (depending on city size)), it covers 45.0% of U.S. and 21.1% of global NO_x emissions in our datasets. Our results empirically validate ~~long-standing~~ 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. Our model inputs are sampled at the single TROPOMI pixel closest to each plant, so the analysis characterizes NO₂ enhancements at that pixel rather than plume-wide structure. 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~~ Several additional variables that plausibly influence detectability ~~were unavailable from TROPOMI or ERA5, including~~ could be considered in future work, such as vertical wind shear ~~;(derivable from ERA5 winds at multiple pressure levels)~~, boundary-layer mixing proxies from ERA5, and aerosol optical depth ~~;and more detailed measures of turbulence~~ (see Sect. 2.4), which would require collocation with a separate product such as the Copernicus Atmosphere Monitoring Service (CAMS) reanalysis (Inness et al., 2019). While our quantitative results apply to TROPOMI-like pixel sizes (3.5–7 km), the qualitative feature dependences follow from instrument-agnostic physics and should generalize to finer-resolution NO₂ sensors such as GOSAT-GW (Japan Aerospace Exploration Agency, 2024), though absolute detectability would shift. Finally, while this study predicts whether plumes are detectable, the next critical step is to use the same ~~environmental and observational predictors~~ meteorological, environmental, sensor, and power-plant variables 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.

785 *Code and data availability.* Data used in the paper are available on Zenodo (<https://doi.org/10.5281/zenodo.17706582>). Code will be released on GitHub upon acceptance.

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

Table A1. Annual TROPOMI ~~snapshot-observation~~ filtering for U.S. ~~top-500~~ Top 500 NO_x emitters (2019–2024). ~~The filtering removes snapshots where~~ Retained counts reflect the modeling subset, i.e., observations with no missing input features, a paired CAMPD record, and a target ~~plants are within~~ plant outside interference zones (Sect. 3.3). Dataset: full quality-filtered U.S. dataset ($n = 666,222$ TROPOMI observations); “Retained” columns sum to the U.S. modeling subset of $n = 189,713$ observations from ~~nearby urban areas or other major power~~ 171 plants.

Year	Unfiltered	Retained	Removed	Removal Rate
2019	84104,658 <u>672</u>	3330,453 <u>966</u>	5173,205 <u>706</u>	60.570.4%
2020	82108,795 <u>786</u>	3230,517 <u>838</u>	5077,278 <u>948</u>	60.771.7%
2021	82108,063 <u>053</u>	3230,023 <u>934</u>	5077,040 <u>119</u>	61.071.4%
2022	86115,676 <u>294</u>	3332,723 <u>808</u>	5282,953 <u>486</u>	61.171.5%
2023	82112,745 <u>427</u>	3231,631 <u>534</u>	5080,114 <u>893</u>	60.672.0%
2024	85116,228 <u>990</u>	3332,991 <u>633</u>	5184,237 <u>357</u>	60.172.1%
Total	504666,165 <u>222</u>	198189,338 <u>713</u>	305476,827 <u>509</u>	60.771.5%

The filtering statistics show remarkable temporal consistency, with removal rates ~~stable at 60–61~~ ranging 70–72% across all years. This ~~consistency reflects the relatively~~ stability reflects the static nature of interference zones: urban boundaries and power plant configurations change slowly compared to the annual observation cycle. The retained ~~198,338 snapshots~~ (39.3189,713 observations (28.5% of the original dataset) ~~represent observations are those~~ 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: Per-Country Statistics for Duplicate Filtering in the CoCO₂ Catalog

Of the 16,461 candidate facilities in the CoCO₂ catalog, our duplicate filter (plants within 3 km sharing identical reported emissions across CO₂, NO_x, SO_x, CO, and CH₄) removed 3,320 entries (20%). The geographic distribution is highly non-uniform:

- China: 1,696 removed (45% of Chinese facilities in the catalog)
- India: 498 removed (48%)
- Russia: 163 removed (25%)
- Japan: 160 removed (13%)
- Indonesia: 101 removed (39%)

- United States: 1 removed (0.035% of 2,850 facilities)

China and India together account for 66% of all removed entries (2,194 of 3,320). The minimal removal rate in the United States is consistent with the more rigorous facility-ID system maintained by the U.S. EPA. The high removal rates in China and India likely reflect catalog integration challenges, where the same physical facility may appear under multiple identifiers (different agencies, naming conventions, or transliterations); these duplicate entries are not necessarily indicative of errors in the underlying emission estimates, but they do require de-duplication for our source-attribution analysis.

Appendix C: 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~~ cross-entropy with log-its loss function. We performed a search over seven learning rates—~~1~~ $1 \times 10^{-5}, 2 \times 10^{-5}, 5 \times 10^{-5}, 1 \times 10^{-4}, 2 \times 10^{-4}, 5 \times 10^{-4}, 1 \times 10^{-3}$ —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~~ 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~~ standard deviation of the results.

Appendix D: Complementary Results for Item Split Experiments

D1 ROC Curves for the U.S. and Global Models

Table 3 reports the Area Under the Curve (AUC) as a threshold-free summary of each model’s discriminative ability. Figure D1 shows the underlying Receiver Operating Characteristic (ROC) curves on the held-out test split for the U.S. (hourly NO_x , “All”) and global (annual NO_x , “All”) models. Each curve traces the trade-off between the true-positive rate (TPR; recall) and the false-positive rate (FPR; false-alarm rate) as the decision threshold sweeps from 0 to 1; the dashed diagonal corresponds to a random classifier ($\text{AUC} = 0.5$). For each region, we plot the curve from the single best model (highest validation AUC) among the five training runs; the run-to-run AUC spread is reported as the mean $\pm$ std in Table 3.

Both curves lie above the diagonal across the full operating range, with U.S. $\text{AUC} = 0.83$ and global $\text{AUC} = 0.80$ for the plotted best-of-five models, consistent with the 5-run means of 0.819 ± 0.006 and 0.801 ± 0.003 in Table 3. The trade-off between recall and false alarms can be read directly off the curves at any chosen threshold. The U.S. and global curves track

each other closely, indicating consistent discriminative behavior despite the different emission-rate distributions and temporal resolutions of the two inventories.

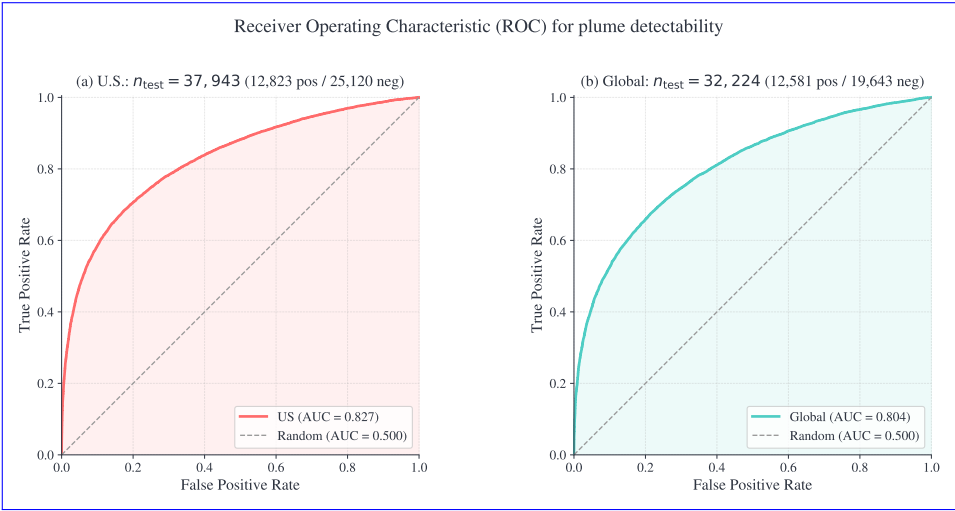


Figure D1. Receiver Operating Characteristic (ROC) curves for the U.S. (hourly NO_x) and global (annual NO_x) MLP detectability models on the “All” subset, evaluated on the held-out test split. For each region, the plotted curve is from the single best model (highest validation AUC) among five training runs; the dashed diagonal is the random-classifier baseline ($\text{AUC} = 0.5$). Dataset: modeling subset (Sect. 3.3; 171 U.S. plants, $n = 189,713$ observations; 1,065 global plants, $n = 161,118$ observations); the held-out test split contains $n = 37,943$ U.S. and $n = 32,224$ global observations.

835 **D2 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; both training and the subsequent analysis are restricted to these 171 plants.

SHAP Score Analysis The SHAP analysis confirms the permutation results—, with NO_x emissions dominate globally—dominating in both the U.S. and global datasets, while adding plant-level nuance. We map Figure D2 maps the top feature per plant (Fig. D2a,e) in panels (a) and (c), and the regional modal top feature on a grid (Fig. D2b, in panels (b) and (d)). Hourly NO_x emissions are the top feature for 151 of 171 U.S. plants (annual NO_x for 805 of 1,065 global plants); surface altitude, surface classification, and wind speed appear as top features for the remaining U.S. plants. Globally, the secondary top features include the 440 nm surface albedo in the NO_2 window (124 plants), the 758 nm surface albedo (25), wind speed (22), surface altitude (15), primary fuel type (14), 2 m temperature (9), surface pressure (9), and TOA incident solar radiation (8).

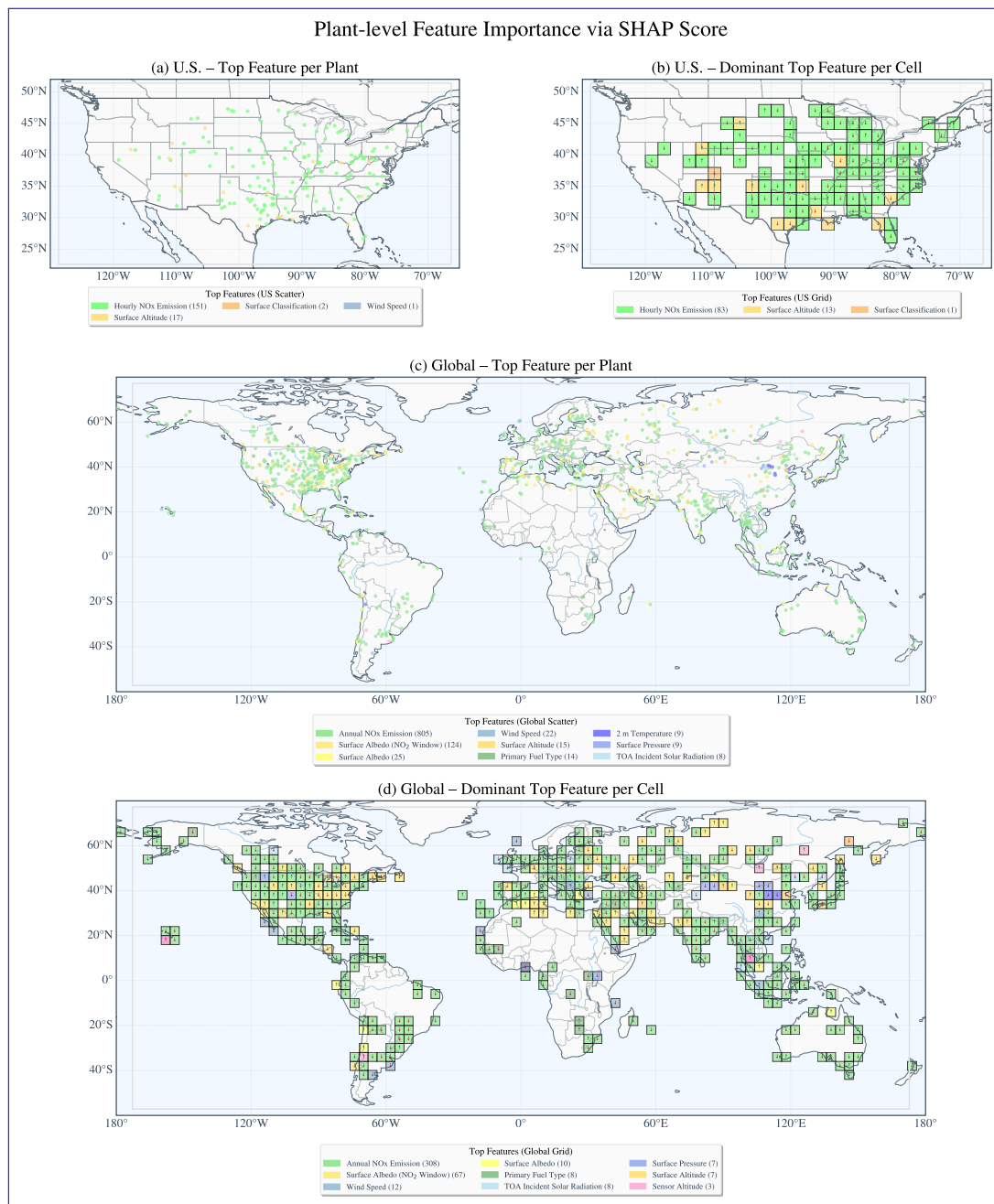


Figure D2. Dominant predictive features of plume detectability across the UPlant-level feature importance via SHAP score. S. and globally (predictors include nearby-city and nearby-plant statistics). **(a,c)** Plant-level scatter: each point is a power plant colored by its most influential feature. **(b,d)** Grid-mode maps: within each cell (2° U.S.; 4° global), the color marks the modal top feature; arrows indicate positive. Legend counts (↑ in parentheses) report the number of plants (a, c) or negative-grid cells (↓, d) effects for which that feature ranks first; the legend lists the top 9 features, so a small number of plants whose top feature falls outside this list are plotted but not enumerated. Colors SHAP values are consistent across panels computed on the held-out test split of the dataset. Dataset: modeling subset (Sect. 3.3; 171 U.S. plants, $n = 189,713$ observations; 1,065 global plants, $n = 161,118$ observations); the held-out test split contains $n = 37,943$ U.S. and $n = 32,224$ global observations.

D3 Correlation between Per-Plant Model Performance and Key Features

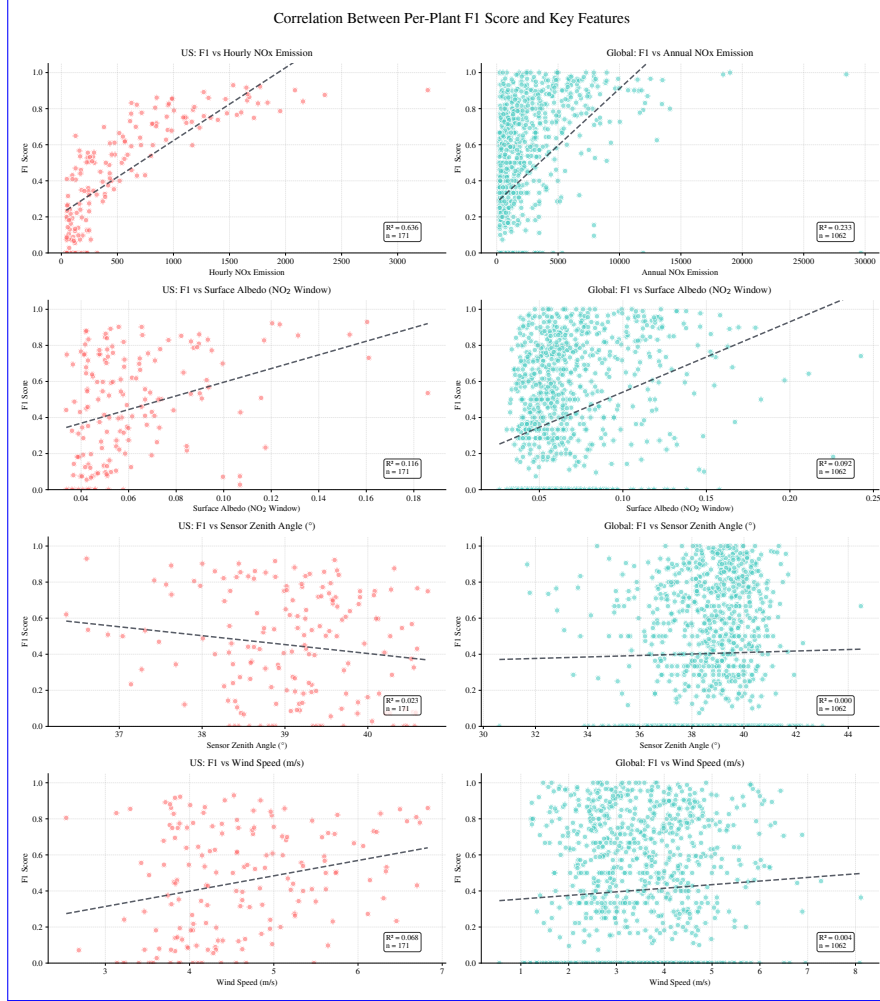


Figure D3. Correlation between F1 Score and Key Features for U.S. and Global Power Plants. Scatter plots show the relationship between the plume detection F1 score and four key features, separated for the U.S. dataset (left column, $n=171$, $n = 171$ plants) and a Global dataset (right column, $n=1062$, $n = 1,062$ plants). F1 scores per plant are computed as in Fig. 8. We filtered-out power further restrict to plants with less-than-at least 3 snapshots positive observations, and that's which is why the available power plants for global analysis drop from 1065 to 1,062. The features (aggregated as plant-level means) are, from top to bottom: NO_x emissions, surface albedo (NO₂ window), 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 (hourly for the U.S., annual for global) in both datasets, especially in the U.S. ($R^2=0.670$, $R^2 = 0.686$). Dataset: modeling subset (Sect. 3.3; 171 U.S. plants, $n = 189,713$ observations; 1,065 global plants, $n = 161,118$ observations), further restricted to plants with at least 3 positive observations (171 U.S./1,062 global).

Appendix E: Cross-pipeline consistency on overlapping plants

850 As an additional check on the U.S. and global pipelines, we compared their per-plant detectability estimates on the 160 plants present in both runs (matched within < 10 m). Per-plant detectabilities agree closely (Pearson $r = 0.96$). The small residual offset (mean $+0.036$; U.S. mean 0.163 vs. global 0.199) likely reflects the temporal gap between the two samples: the global run is anchored to 2018, while the U.S. run spans 2019–2024, a period over which U.S. NO_x emissions have continued to decline.

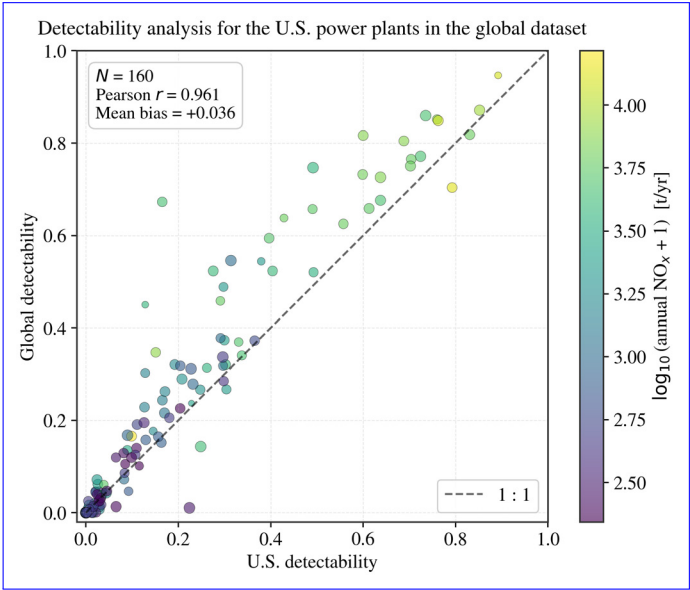


Figure E1. Two-dimensional probability-of-detection (POD) analysis for TROPOMI NO_2 plume detection from power plants. Heatmaps show POD as a function of NO_x emission rate and key environmental variables for Per-plant detectability comparison between the U.S. facilities pipeline (left column; hourly emission in kg/h 2019–2024) and the global facilities pipeline (right column; annual emission 2018) for the 160 plants present in both runs (matched within < 10 m). Top row: NO_x emission vs. surface albedo in Each point is a plant, colored by $\log_{10}$ of the NOCoCO_2 spectral window. Bottom row: annual NO_x emission vs. 10-m wind speed. Color intensity indicates POD emissions (red = low in t yr^{-1} , green = high with a $+1$ offset for plants near zero). The dashed diagonal marks 1:1 agreement. Pearson $r = 0.961$; cell annotations display POD values and sample counts (n) mean bias $= +0.036$. Binning uses a quantile-based strategy Dataset: full quality-filtered dataset (1st–99th percentile) to ensure balanced feature distribution $n = 666,222$ U.S. and avoid extreme values $n = 875,686$ global TROPOMI observations).

Appendix F: Power Plant Split Experiments

855 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 Sect. 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~~ 171 power plants lie outside interference zones across every analyzed year. We retained ~~183~~ 171 from Top 500 U.S. plants (~~52~~ 50 from Top 100, ~~32~~ 30 from Top 50, 13 from Top 20) and ~~1,065 from Top 6,035 from Top 6000~~ 1,065 from Top 6,035 from Top 6000 global plants (35 from Top 100, 18 from Top 50) ~~with ≥ 1 year of usable data~~ , 5 from Top 20). 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 F1 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~~ changes in performance for comprehensive datasets. In the U.S., accuracy ~~drops only 2–7% for the “All” dataset~~ (0.750 on the “All” dataset is essentially preserved (0.757 $\rightarrow$ 0.697 0.768 for annual NO_x, 0.763 0.775 $\rightarrow$ 0.742 0.779 for hourly NO_x), while global accuracy declines 6.7% (0.745 by about 5.6 percentage points (0.746 $\rightarrow$ 0.695 0.690)). AUC scores remain strong across all scenarios, indicating the model maintains good discriminative ability when tested on entirely new facilities for the “All” models, with larger declines on the global high-emitter subsets.

~~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.~~ On the high-emitter subsets, F1 actually rises in all three configurations, while Cohen’s κ generally declines, reflecting the higher positive-class prevalence in those subsets and the small per-subset test set.

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~~ plants, particularly for high-emitting ~~subsets—the subsets, the~~ maintained discriminative ability across diverse facilities validates the robustness of our approach for real-world deployment scenarios.

Appendix G: Feature Importance and Correlation

Figure G1 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.

Table F1. Multi-Layer Perceptron ~~Performance-performance~~ (~~Power-Plant-Split~~~~power plant split~~) —for U.S. and ~~Global~~~~global~~ datasets. Dataset: modeling subset (Sect. 3.3: 171 U.S. plants, $n = 189,713$ observations; 1,065 global plants, $n = 161,118$ observations); each “Top X ” row uses the corresponding subset of Table 2.

Region	Emissions	Dataset	Accuracy	Precision	Recall	F1 Score	AUC
U.S.	Annual NO _x	All	0.697 ± 0.003 0.768 ± 0.004	0.569 ± 0.005 0.640 ± 0.021	0.710 ± 0.010 0.446 ± 0.021	0.632 ± 0.002 0.525 ± 0.008	0.772 ± 0.002 0.747 ± 0.003
		Top-100 Top 100	0.626 ± 0.004 0.670 ± 0.002	0.721 ± 0.006 0.690 ± 0.009	0.522 ± 0.010 0.716 ± 0.021	0.605 ± 0.011 0.702 ± 0.006	0.657 ± 0.003 0.720 ± 0.004
		Top-50 Top 50	0.675 ± 0.007 0.725 ± 0.006	0.745 ± 0.009 0.801 ± 0.005	0.708 ± 0.020 0.793 ± 0.014	0.726 ± 0.012 0.797 ± 0.006	0.721 ± 0.005 0.768 ± 0.006
		Top 20	0.722 ± 0.005	0.797 ± 0.017	0.827 ± 0.037	0.811 ± 0.009	0.732 ± 0.010
	Hourly NO _x	All	0.742 ± 0.003 0.779 ± 0.005	0.631 ± 0.006 0.667 ± 0.015	0.712 ± 0.005 0.644 ± 0.016	0.669 ± 0.001 0.655 ± 0.003	0.801 ± 0.002 0.813 ± 0.003
		Top-100 Top 100	0.657 ± 0.005 0.742 ± 0.002	0.756 ± 0.012 0.805 ± 0.006	0.557 ± 0.031 0.750 ± 0.012	0.640 ± 0.016 0.776 ± 0.004	0.708 ± 0.013 0.811 ± 0.004
		Top-50 Top 50	0.717 ± 0.006 0.777 ± 0.004	0.769 ± 0.015 0.852 ± 0.008	0.764 ± 0.017 0.813 ± 0.012	0.766 ± 0.002 0.832 ± 0.004	0.769 ± 0.007 0.832 ± 0.004
		Top 20	0.761 ± 0.011	0.816 ± 0.008	0.863 ± 0.021	0.839 ± 0.009	0.783 ± 0.010
Global	Annual NO _x	All	0.695 ± 0.005 0.690 ± 0.006	0.596 ± 0.011 0.590 ± 0.015	0.538 ± 0.018 0.534 ± 0.021	0.565 ± 0.007 0.560 ± 0.006	0.730 ± 0.004 0.725 ± 0.005
		Top-100 Top 100	0.588 ± 0.020 0.616 ± 0.029	0.851 ± 0.015 0.836 ± 0.016	0.583 ± 0.034 0.644 ± 0.042	0.691 ± 0.022 0.727 ± 0.027	0.623 ± 0.017 0.614 ± 0.018
		Top-50 Top 50	0.616 ± 0.044 0.679 ± 0.109	0.918 ± 0.010 0.909 ± 0.011	0.623 ± 0.063 0.715 ± 0.143	0.740 ± 0.030 0.793 ± 0.077	0.646 ± 0.032 0.627 ± 0.033
		Top 20	0.959 ± 0.043	0.992 ± 0.003	0.966 ± 0.045	0.978 ± 0.023	0.860 ± 0.020

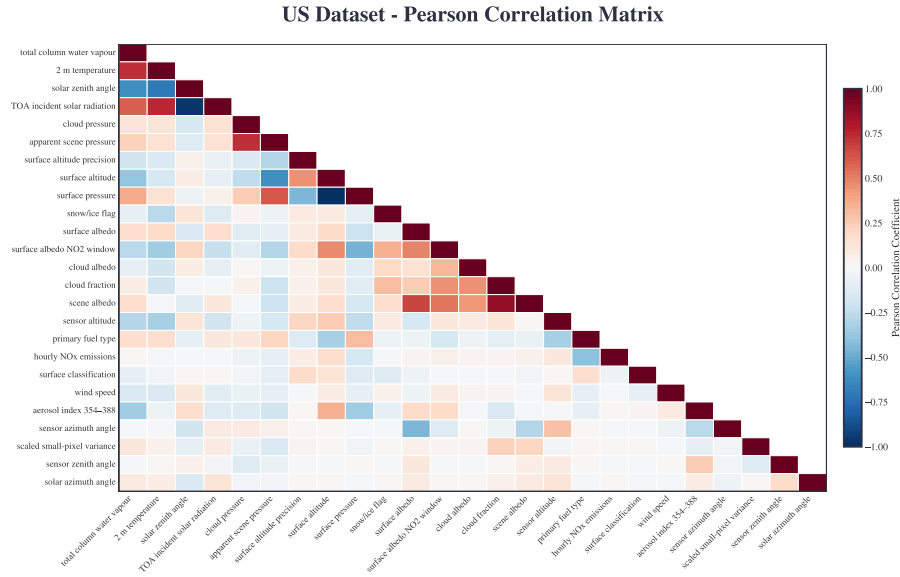
885 First, atmospheric variables such as surface altitude and surface pressure exhibit an almost perfect negative correlation ($\rho = -0.994$ in the U.S., $\rho = -0.986$ globally). This simply mirrors the barometric law: as you ascend, the weight of the air above you diminishes, driving pressure down.

Likewise, solar zenith angle and ~~total incoming solar radiation (TISR)~~~~TOA incident solar radiation~~ 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
890 a horizontal surface.

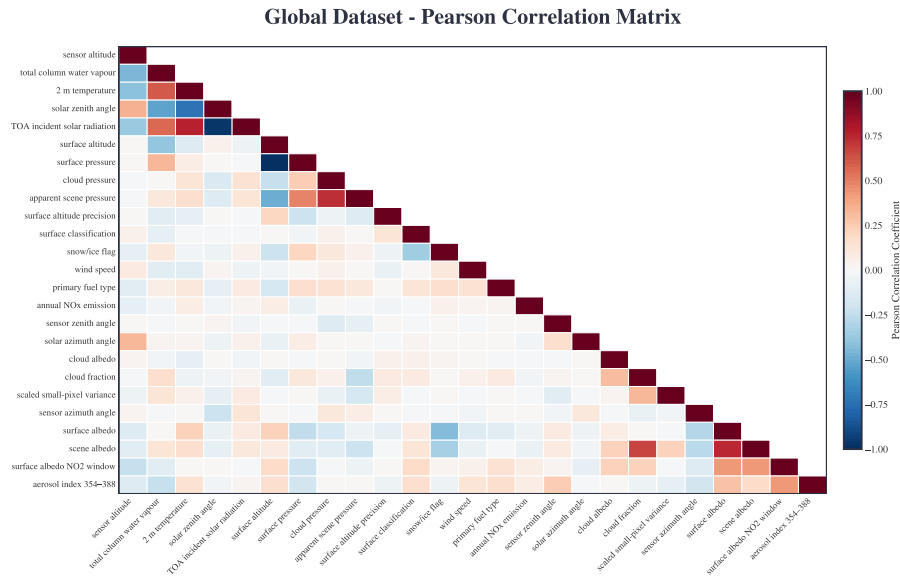
One additional U.S.-~~specific-specific~~ pattern links cloud cover to surface brightness: cloud fraction and scene albedo correlate at ~~$\rho = 0.865$~~ ~~$\rho = 0.863$~~ , as more cloud cover generally leads to brighter reflectance in visible wavelengths.

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 multicol-linearity in regression mod-
895 els. Future work can mitigate this risk by dropping one feature from each pair or by applying ~~dimensionality-reduction~~~~dimensionality-reduction~~ techniques (e.g., principal component analysis) to extract orthogonal components.

Table G1 records the permutation importance across four model configurations we displayed in ~~Section ??~~~~Sect. 4.4~~. The bar chart visualization of Table G1 is Figure 9.



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



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

Figure G1. 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).

[Dataset: modeling subset \(Sect. 3.3: 171 U.S. plants, \$n = 189,713\$ observations; 1,065 global plants, \$n = 161,118\$ observations\).](#)

Feature	U.S.	Global
<i>Sensor Features (TROPOMI)</i>		
sensor zenith angle	0.0232 0.0295	0.0384 0.0327
sensor azimuth angle	0.0038 0.0029	0.0071 0.0068
sensor altitude	0.0046 0.0089	0.0339 0.0344
scaled small-pixel variance	0.0034 0.0036	0.0050 0.0036
<i>Power Plant Features</i>		
annual NO _x emissions	—	0.1316 0.1351
hourly NO _x emissions	0.1934 0.2226	—
primary fuel type	0.0048 0.0050	0.0521 0.0493
<i>Meteorology Features</i>		
<i>TROPOMI</i>		
10 m wind speed 0.0280 0.0250-cloud albedo	0.00035 0.0006	0.00024 0.0005
cloud pressure	0.0032 0.0024	0.0039 0.0041
cloud fraction	0.0029 0.0032	0.0097 0.0090
surface pressure	0.0030 0.0039	0.0209 0.0276
solar zenith angle	0.0025 0.0026	0.0169 0.0177
solar azimuth angle	0.0014 0.0009	0.0088 0.0095
apparent scene pressure	0.0038 0.0031	0.0130 0.0110
aerosol index 354–388	0.0028	0.0088 0.0086
<i>ERA5</i>		
<u>100 m wind speed</u>	<u>0.0291</u>	<u>0.0226</u>
2 m temperature	0.0049 0.0034	0.0199 0.0209
total column water vapour	0.0034 0.0086	0.0095 0.0084
TOA incident solar radiation	0.0059 0.0046	0.0218 0.0233
<i>Environment Features</i>		
<i>TROPOMI</i>		
surface classification	0.0077 0.0095	0.0203 0.0211
snow/ice flag	0.0021 0.0023	0.0087 0.0082
scene albedo	0.0027 0.0017	0.0134 0.0137
surface albedo	0.0030 0.0040	0.0290 0.0315
surface albedo (NO ₂ window)	0.0076 0.0092	0.0574 0.0549
surface altitude	0.0108 0.0222	0.0389 0.0476
surface altitude precision	0.0024 0.0056	0.0222 0.0257

Table G1. Permutation importance for the U.S. “All” Hourly NO_x model and the Global models “All” Annual NO_x model (item-split), computed on the held-out test split as described in Sect. 3.5. “—” marks a feature that is absent. Values rounded to 4 decimals. Dataset: modeling subset (Sect. 3.3; 171 U.S. plants, $n = 189,713$ observations; 1,065 global plants, $n = 161,118$ observations); the held-out test split contains $n = 37,943$ U.S. and $n = 32,224$ global observations.

G1 Detectability as a Function of Two Features

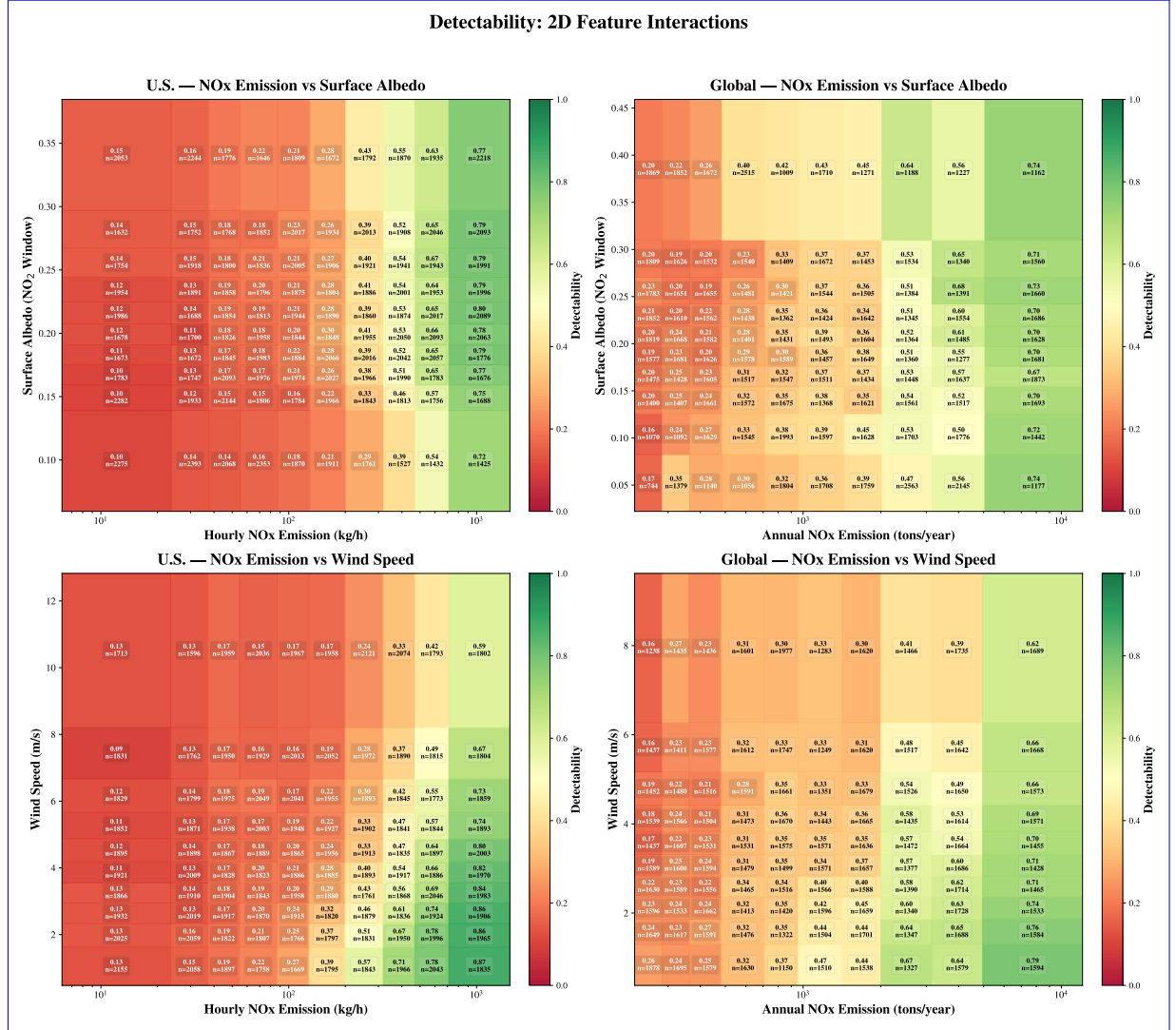


Figure G2. Two-dimensional detectability analysis for TROPOMI NO₂ plume detection from power plants. Heatmaps show empirical detectability as a function of NO_x emission rate and key environmental variables for U.S. facilities (left column; hourly emission in kg h⁻¹) and global facilities (right column; annual emission in t yr⁻¹). Top row: NO_x emission vs. surface albedo in the NO₂ spectral window. Bottom row: NO_x emission vs. 100 m wind speed. Color intensity indicates detectability (red = low, green = high); cell annotations display detectability values and sample counts (*n*). Binning uses a quantile-based strategy (1st–99th percentile) to ensure balanced feature distribution and avoid extreme values. Dataset: modeling subset (Sect. 3.3; 171 U.S. plants, *n* = 189,713 observations; 1,065 global plants, *n* = 161,118 observations).

Although cloud fraction has low permutation importance in Table G1, this reflects the upstream $qa > 0.75$ filter rather than physical irrelevance. After the qa filter, cloud fraction is truncated to values below ~ 0.3 (Fig. G3a), and within this surviving range detection probability still decreases monotonically with cloud fraction, from 0.41 to 0.29 in the U.S. (Fig. G3b) and from 0.45 to 0.35 globally (Fig. G3c). The cloud-fraction effect is therefore absorbed by the qa filter upstream.

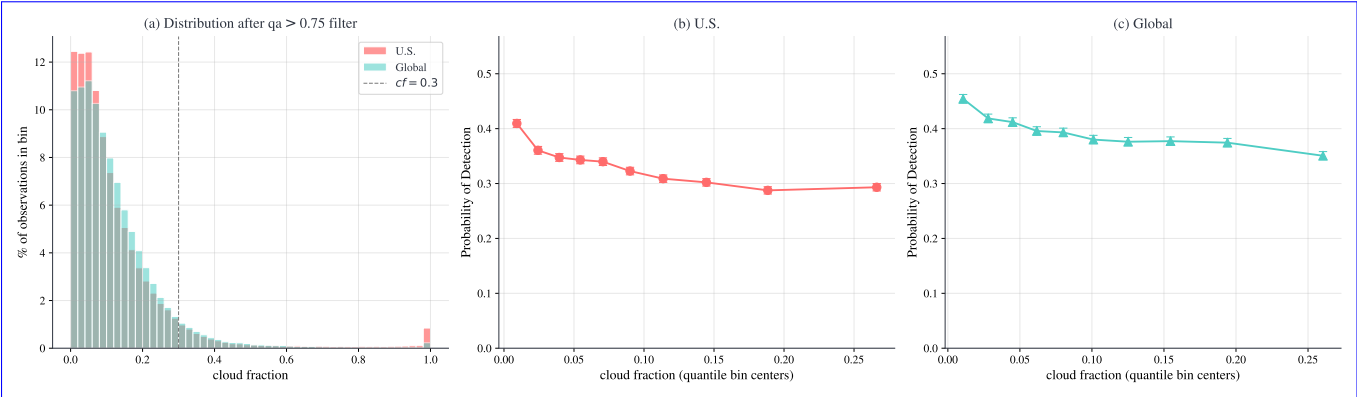


Figure G3. Detectability as a function of cloud fraction, restricted to observations passing the $qa > 0.75$ filter. (a) Distribution of cloud fraction in the U.S. (red) and global (teal) samples; the qa filter truncates cloud fraction to values below ~ 0.3 . (b) U.S. probability of detection versus cloud fraction, in quantile bins. (c) Same as (b) for the global sample. Dataset: modeling subset (Sect. 3.3; 171 U.S. plants, $n = 189,713$ observations; 1,065 global plants, $n = 161,118$ observations).

Figure G4 shows the TROPOMI pixel area as a function of sensor zenith angle (VZA). The 25 km² flagged-area threshold (Sect. 3.2, Step 5) is crossed at $VZA \approx 26^\circ$: below this angle a single pixel is smaller than 25 km² and a detection requires at least two contiguous flagged pixels, while above it a single flagged pixel already suffices. This pixel-quantization step contributes to the rising flank of the VZA–detectability curve discussed in Sect. 4.4.

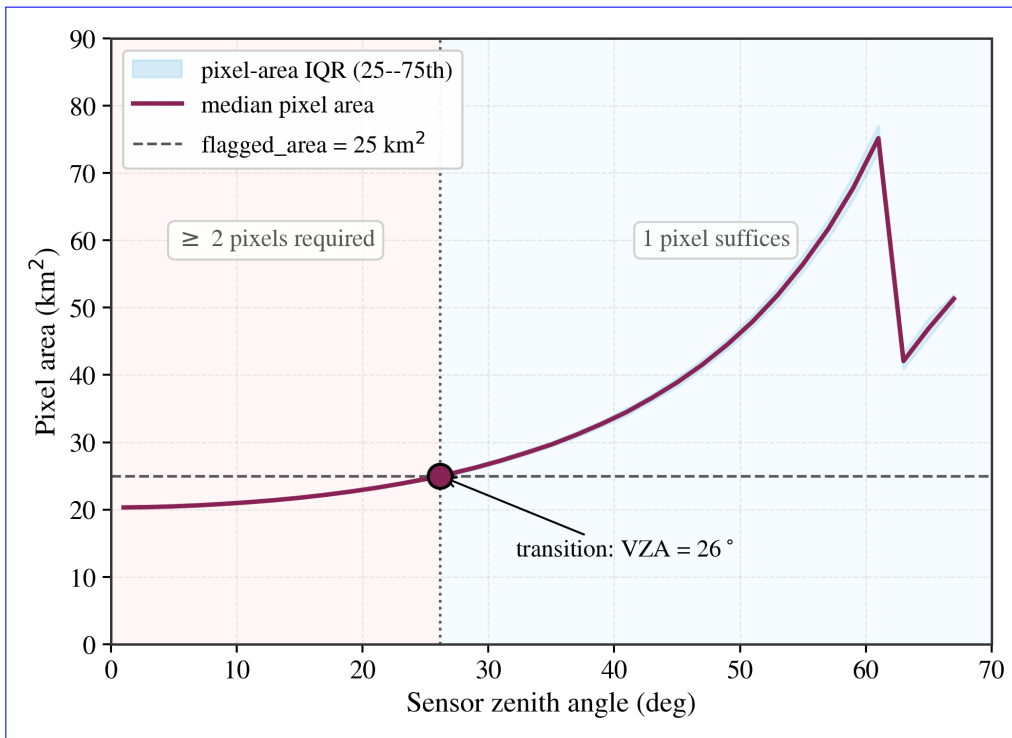


Figure G4. TROPOMI pixel area versus sensor zenith angle. The solid line is the median pixel area, with the shaded band showing the 25th–75th percentile range. The dashed horizontal line marks the 25 km² flagged-area threshold used in our plume-detection algorithm. Below VZA $\approx 26^\circ$, a single pixel is smaller than 25 km², so detection requires at least two contiguous flagged pixels; above 26°, one pixel already exceeds the threshold.

910 Appendix H: Case Studies in Automated Plume Detection Algorithm Performance

To assess the performance of our Automated Plume Detection Algorithm (Section Sect. 3.2), 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.

present representative examples across the four classification outcomes. A successful classification results in either a **True Positive** yields either a true positive (TP), where a genuine plume is correctly identified and attributed, or a **True Negative** true negative (TN), where the absence of a plume from the target plant is correctly noted. Conversely, misclassifications Misclassifications fall into two categories:

False Positives: The algorithm incorrectly attributes emissions from other sources a false positive (FP), where the algorithm wrongly attributes an enhancement to the target power plant. **False Negatives:** The algorithm fails to detect a plant, and a false negative (FN), where it misses 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:

(i) **Interference**: Signals, where signals from nearby industrial facilities or urban centers are confused with the target plume. Conversely, or where the interference mask itself can sometimes obscure a genuine plume, obscures a genuine target plume; (ii) **Incorrect Threshold**: The detection, where the threshold is either too high, causing (faint but real plumes to be missed, are missed) or too low, causing background noise to be (background noise is flagged as a plume.); and (iii) **Regional Gradients**: Natural, where natural variations in background NO₂ levels, such as the sharp contrast between land and sea, (e.g., land-sea contrasts or terrain-driven gradients) 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 TROPOMI NO₂ concentration field around the target plant, with the target facility (red cross), nearby industrial sources (blue triangles), wind direction (blue triangle and arrow), and urban centers (yellow dots) marked. The top-middle panel overlays the detected plume contour (red outline) overlaid; nearby cities and other power plants and their interference-zone masks are also shown where relevant. The right panel shows ArcGIS World Imagery for the same area for geographic context (Esri, n.d.). The header of each case lists the classification outcome, plant identifier and country, observation date, annual NO_x emissions, wind direction and speed, 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, detected plume area (when applicable).

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

H1 Correct Classifications

H2 Successful Detections

H1.1 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 ?? shows an ideal detection, while Figure ?? demonstrates a successful detection of a weaker, more diffuse plume.

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.

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.

H1.1 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.

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.

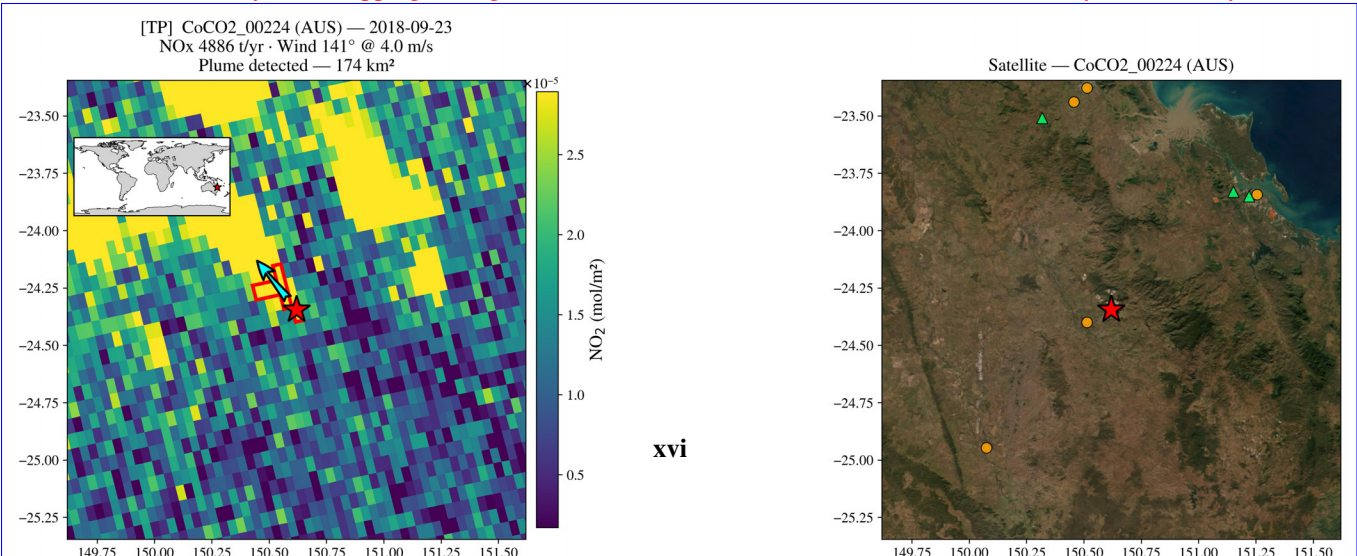
A correct true negative. Although a plume is visible, most likely it originates from other emission sources, and the interference zone correctly prevents it from being attributed to the power plant under study. Satellite Image Source: Esri, i-cubed, USDA, USGS, AEX, GeoEye, Getmapping, Aerogrid, IGN,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

H2 Failed Detections and Their Causes

H1.1 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.

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.



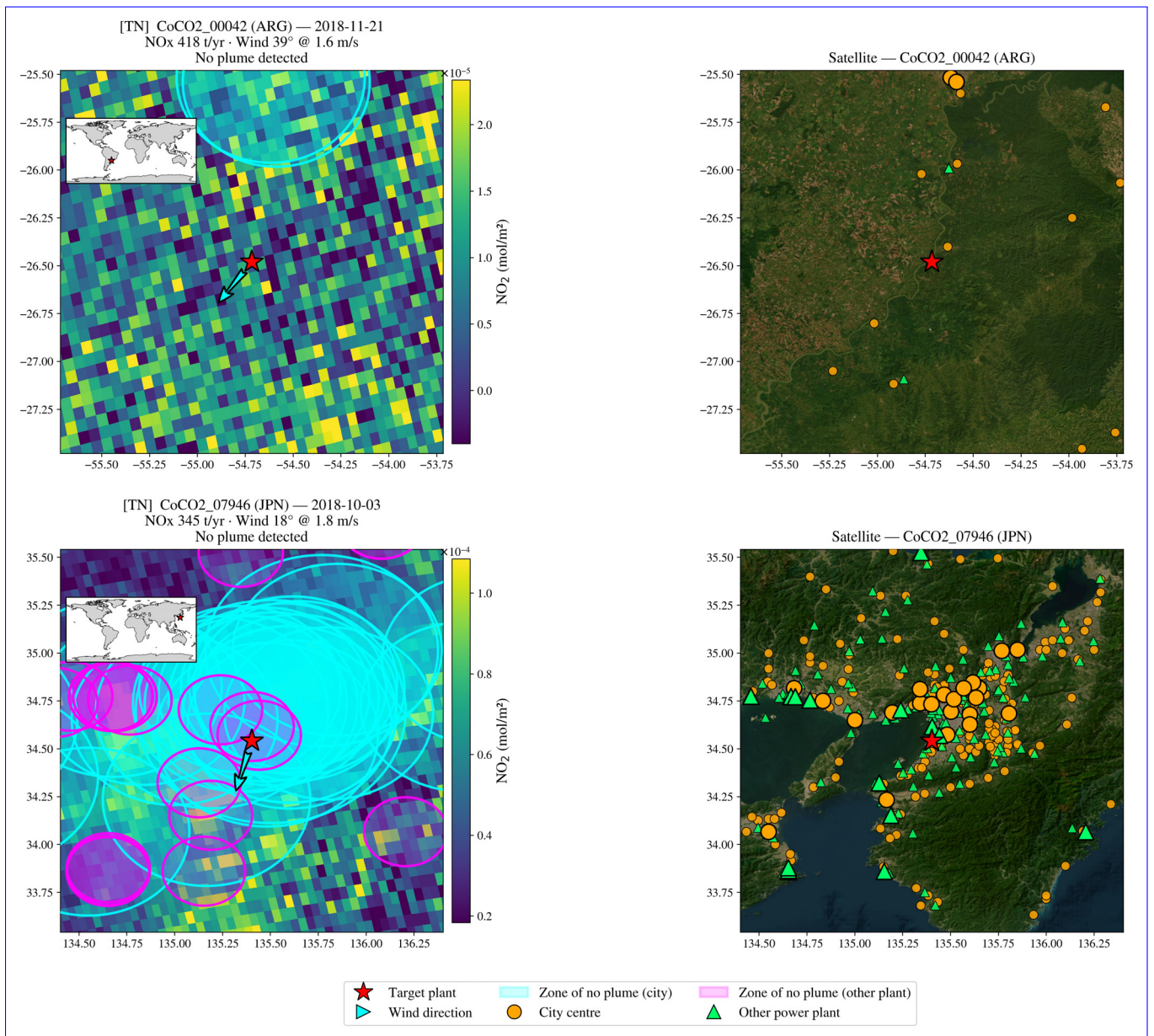


Figure H2. Cause: Interference Mask (False Negative) Two true-negative cases. A prominent The algorithm correctly returns no detection. The top example shows a quiescent scene with no discernible plume, while the bottom example shows a scene with visible NO₂ enhancements that originate from sources other than 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 zones (cyan and pink shaded areas) correctly prevent attribution to avoid confusion with other sources the target. Satellite Image Source Basemap: Esri, i-cubed, USDA, USGS, AEX, GeoEye, Getmapping, Aerogrid, IGN, IGP, UPR-EGP, and the GIS User Community + Powered by Esri ArcGIS World Imagery (Esri, n.d.).

The detection outcome is highly sensitive to the chosen NO_2 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).

H2 Misclassifications and Their Causes

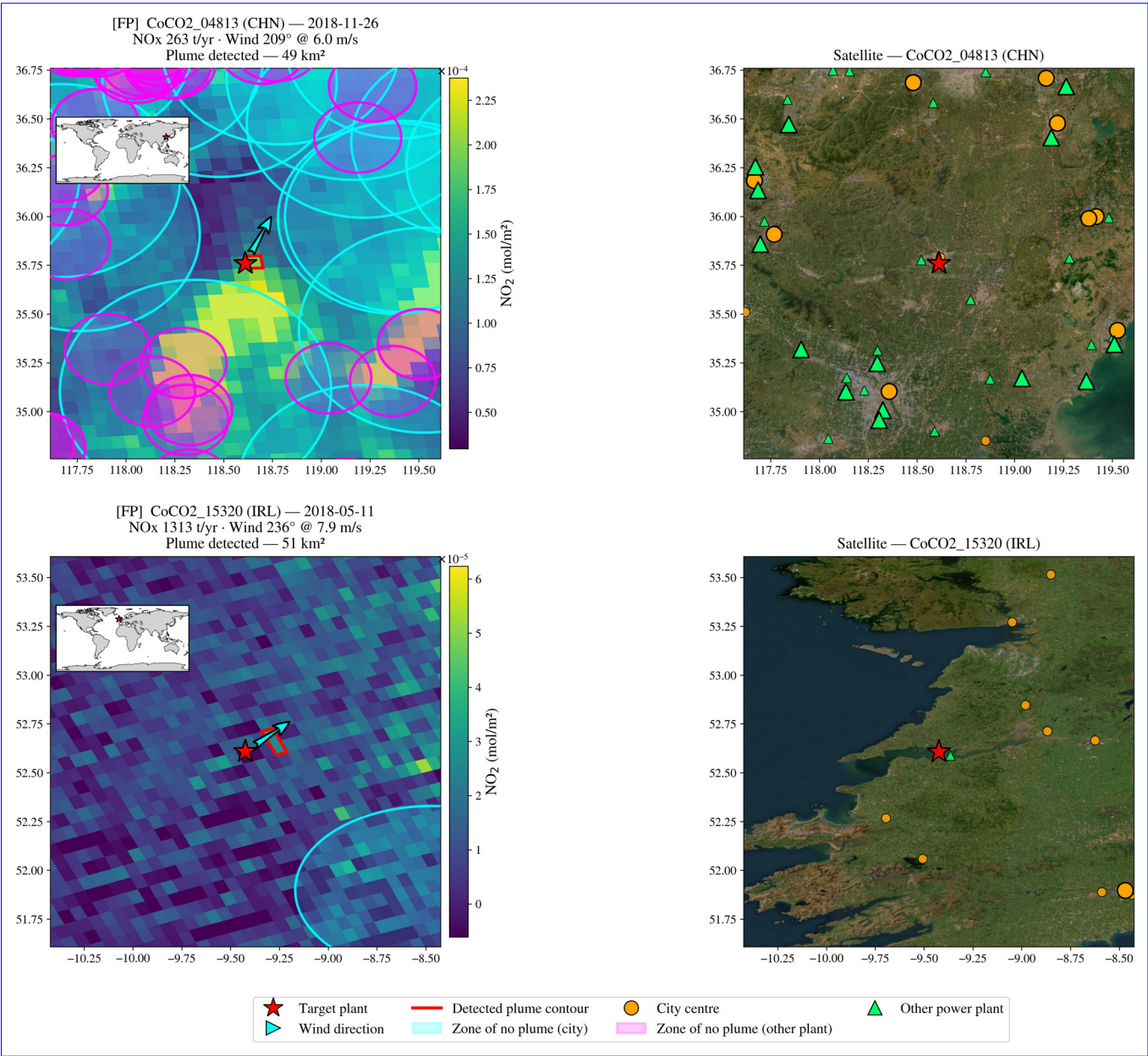


Figure H3. Cause: High Threshold (False Negative) Two false-positive cases (part 1 of 2). A real, low-intensity plume is present but falls below the algorithm's detection threshold. The top example shows interference from a nearby urban area being mistaken for a plume. The bottom example shows the detection threshold being set too high to capture such weak signals, so background variability is flagged as a plume. **Satellite Image Source:** Basemap: Esri, i-cubed, USDA, USGS, AEX, GeoEye, Getmapping, Aerogrid, IGN, IGP, UPR-EGP, and the GIS User Community | Powered by Esri ArcGIS World Imagery (Esri, n.d.).

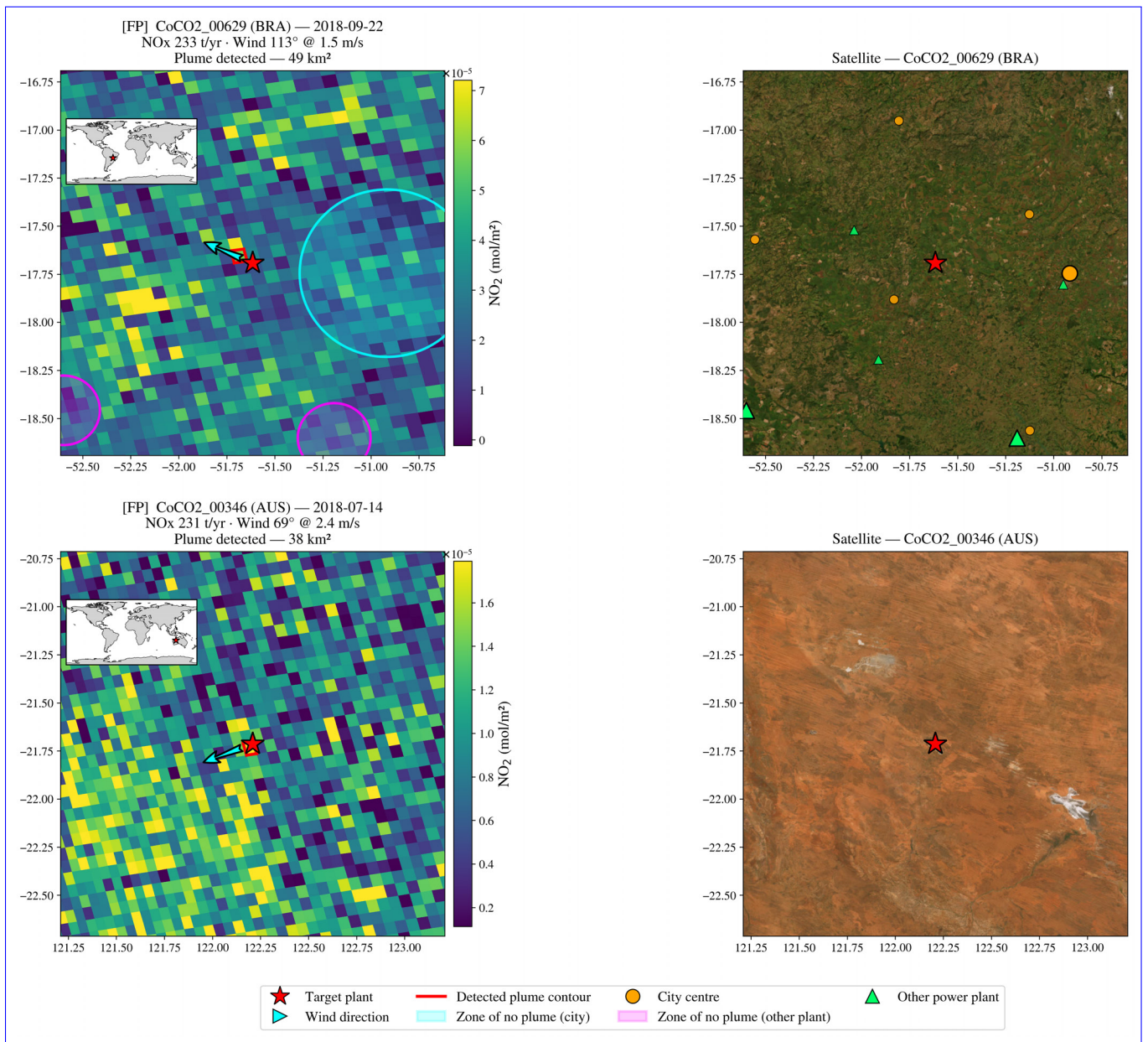


Figure H4. Cause: Low-Threshold (False-Positive) Two false-positive cases (part 2 of 2). The detection threshold is set too low, causing the algorithm to identify Both examples show regional gradients driven by terrain being misinterpreted as a plume where none exists. The flagged area is simply part of the natural background variability in top example shows a sharp NO_2 concentration. Satellite Image Source: Esri contrast against a mountain range, i-cubed, USDA, USGS, AEX, GeoEye, Getmapping, Aerogrid, IGN, IGP, UPR-EGP, and the GIS User Community+ Powered by Esri bottom example shows a similar terrain-driven gradient producing a spurious detection in an arid region. Basemap: ArcGIS World Imagery (Esri, n.d.).

Sometimes, large-scale variations in NO_2 can create sharp gradients that the algorithm misinterprets as a plume. A classic example is the land-sea boundary, where NO_2 levels are often naturally higher over terrestrial areas than over water.

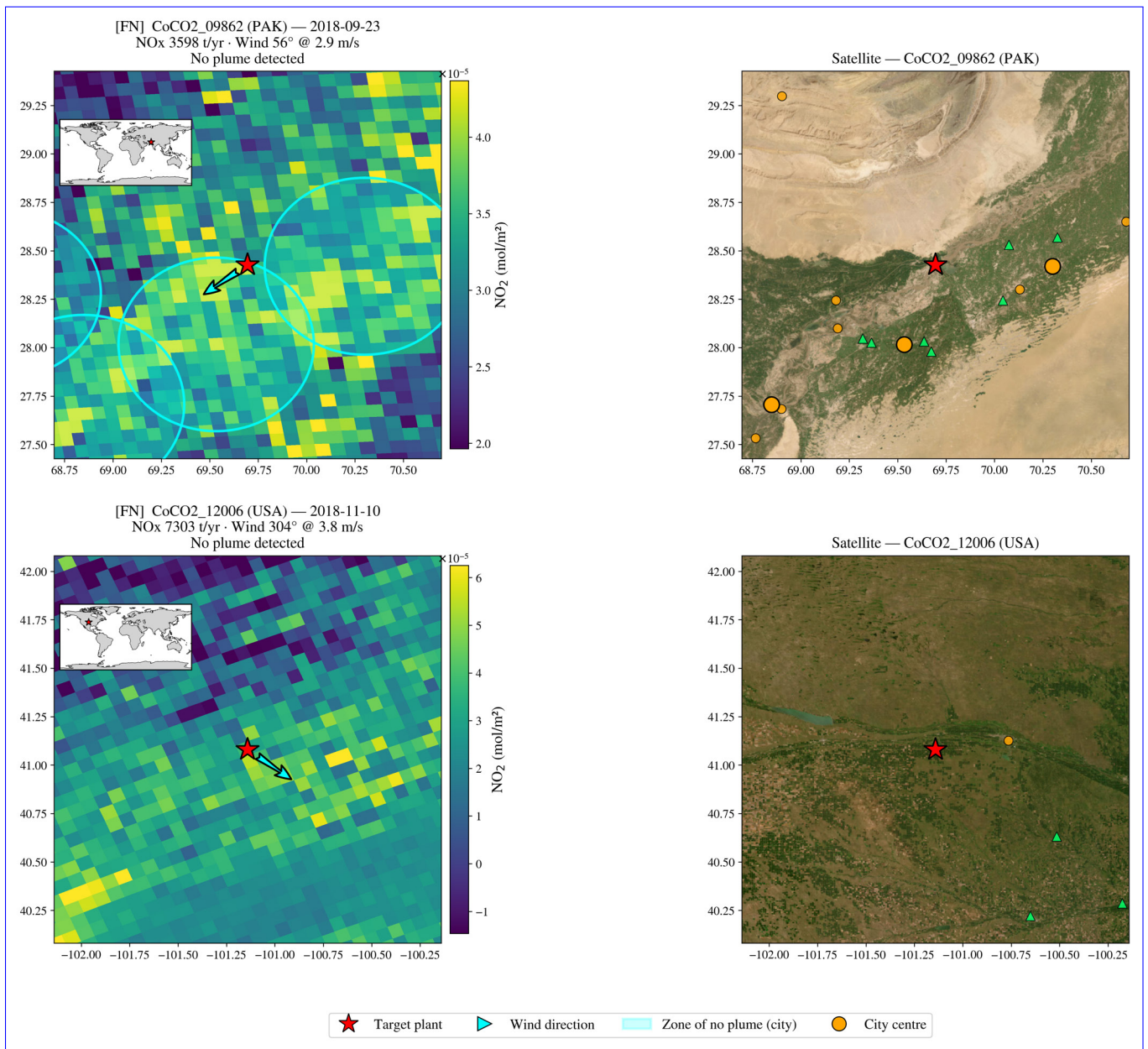


Figure H5. Cause: Topographical Gradient (False Positive) Two false-negative cases. No distinct plume from the target plant that is present not detected because it overlaps with a pre-defined interference mask (cyan shaded area) intended to suppress signals from nearby cities. The algorithm has incorrectly identified the sharp contrast between higher NO₂ concentrations banked against the mountain and lower concentrations in the surrounding area as an emission bottom example shows a real but low-intensity plume that falls below the detection threshold. Satellite-Image-Source: Esri, i-cubed, USDA, USGS, AEX, GeoEye, Getmapping, Aerogrid, IGN, IGP, UPR-EGP, and the GIS User Community | Powered by Esri ArcGIS World Imagery (Esri, n.d.).

Cause: Gradient (False Positive). No distinct plume is present. The algorithm has incorrectly identified the sharp contrast between higher NO_2 concentrations in front of the mountain as plumes. Satellite Image Source: Esri, i-cubed, USDA, USGS, 965 AEX, GeoEye, Getmapping, Aerogrid, IGN, IGP, UPR-EGP, and the GIS User Community | Powered by Esri.

Author contributions. RH performed the experiments and wrote the paper, except for the discussion and conclusion. SW supervised the project and wrote the discussion and conclusion.

Competing interests. The authors declare that they have no competing interests.

Acknowledgements. We sincerely thank Hannah Grauer, Yuhao Nie, [and](#) Dan Klugen for ~~the~~ various discussions. We thank Wenjie Lu, 970 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

- Arribas-Bel, D. and contextily contributors: contextily: Context geo-tiles in Python, <https://github.com/geopandas/contextily>, 2024.
- Beirle, S., Boersma, K. F., Platt, U., Lawrence, M. G., and Wagner, T.: Megacity emissions and lifetimes of nitrogen oxides probed from
975 space, *Science*, 333, 1737–1739, 2011.
- 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.
- Beirle, S., Borger, C., Dörner, S., Eskes, H., Kumar, V., de Laat, A., and Wagner, T.: Catalog of NO_x emissions from point sources as derived from the divergence of the NO₂ flux for TROPOMI, *Earth System Science Data*, 13, 2995–3012, 2021.
- 980 Beirle, S., Borger, C., Jost, A., and Wagner, T.: Improved catalog of NO_x point source emissions (version 2), *Earth System Science Data*, 15, 3051–3073, 2023.
- Boersma, K., Eskes, H., Dirksen, R., Van Der A, R., Veefkind, J., Stammes, P., Huijnen, V., Kleipool, Q., Sneep, M., Claas, J., et al.: An improved tropospheric NO₂ column retrieval algorithm for the Ozone Monitoring Instrument, *Atmospheric Measurement Techniques*, 4, 1905–1928, 2011.
- 985 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.
- Couture, H. D., Alvara, M., Freeman, J., Davitt, A., Koenig, H., Rouzbeh Kargar, A., OConnor, 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.
- 990 Crippa, M., Guizzardi, D., Banja, M., Solazzo, E., Muntean, M., Schaaf, E., Pagani, F., Monforti-Ferrario, F., Olivier, J., Quadrelli, R., et al.: CO₂ emissions of all world countries, JRC Science for Policy Report, European Commission, EUR, 31182, 2022.
- 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.
- 995 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.
- de Foy, B., Lu, Z., Streets, D. G., Lamsal, L. N., and Duncan, B. N.: Estimates of power plant NO_x emissions and lifetimes from OMI NO₂ satellite retrievals, *Atmospheric Environment*, 116, 1–11, 2015.
- 1000 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.
- 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,
1005 2003.
- Esri: World Imagery (basemap), <https://www.arcgis.com/home/item.html?id=10df2279f9684e4a9f6a7f08febac2a9>, arcGIS Online basemap, n.d.

- European Commission: Industrial Emissions Directive (2010/75/EU), <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32010L0075>, 2010.
- 1010 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.
- 1015 Finch, D. P., Palmer, P. I., and Zhang, T.: Automated detection of atmospheric NO₂ plumes from satellite data: a tool to help infer anthropogenic combustion emissions, *Atmospheric Measurement Techniques*, 15, 721–733, 2022.
- 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.
- 1020 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 Gaussian plume model, *Journal of Cleaner Production*, 397, 136 525, 2023.
- 1025 Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., et al.: The ERA5 global reanalysis, *Quarterly journal of the royal meteorological society*, 146, 1999–2049, 2020.
- Inness, A., Ades, M., Agustí-Panareda, A., Barré, J., Benedictow, A., Blechschmidt, A.-M., Dominguez, J. J., Engelen, R., Eskes, H., Fleming, J., et al.: The CAMS reanalysis of atmospheric composition, *Atmospheric Chemistry and Physics*, 19, 3515–3556, 2019.
- 1030 Japan Aerospace Exploration Agency: GOSAT-GW (Global Observing SATellite for Greenhouse gases and Water cycle), <https://www.satnavi.jaxa.jp/files/project/gosat-gw/en/>, accessed: 2026-05-10, 2024.
- Kuhlmann, G., Broquet, G., Marshall, J., Clément, V., Löschner, A., Meijer, Y., and Brunner, D.: Detectability of CO₂ emission plumes of cities and power plants with the Copernicus Anthropogenic CO₂ Monitoring (CO2M) mission, *Atmospheric Measurement Techniques*, 12, 6695–6719, 2019.
- 1035 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.
- Lange, K., Richter, A., and Burrows, J. P.: Variability of nitrogen oxide emission fluxes and lifetimes estimated from Sentinel-5P TROPOMI observations, *Atmospheric Chemistry and Physics*, 22, 2745–2767, 2022.
- 1040 Laughner, J. L. and Cohen, R. C.: Direct observation of changing NO_x lifetime in North American cities, *Science*, 366, 723–727, 2019.
- Lauvaux, T., Giron, C., Mazzolini, M., d’Aspremont, A., Duren, R., Cusworth, D., Shindell, D., and Ciais, P.: Global assessment of oil and gas methane ultra-emitters, *Science*, 375, 557–561, 2022.
- Liu, F., Beirle, S., Zhang, Q., Dörner, S., He, K., and Wagner, T.: NO_x lifetimes and emissions of cities and power plants in polluted background estimated by satellite observations, *Atmospheric Chemistry and Physics*, 16, 5283–5298, 2016.

- 1045 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 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.
- Lorente, A., Folkert Boersma, K., Yu, H., Dörner, S., Hilboll, A., Richter, A., Liu, M., Lamsal, L. N., Barkley, M., De Smedt, I., et al.: Structural uncertainty in air mass factor calculation for NO₂ and HCHO satellite retrievals, *Atmospheric Measurement Techniques*, 10, 759–782, 2017.
- 1050 McDuffie, E. E., Smith, S. J., O'Rourke, P., Tibrewal, K., Venkataraman, C., Marais, E. A., Zheng, B., Crippa, M., Brauer, M., and Martin, 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>, 2020.
- 1055 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.
- Paris Agreement: Paris agreement, in: report of the conference of the parties to the United Nations framework convention on climate change (21st session, 2015: Paris). Retrived December, vol. 4, p. 2, HeinOnline, 2015.
- 1060 Romer, P. S., Duffey, K. C., Wooldridge, P. J., Edgerton, E., Baumann, K., Feiner, P. A., Miller, D. O., Brune, W. H., Koss, A. R., De Gouw, J. A., et al.: Effects of temperature-dependent NO_x emissions on continental ozone production, *Atmospheric Chemistry and Physics*, 18, 2601–2614, 2018.
- Rouet-Leduc, B. and Hulbert, C.: Automatic detection of methane emissions in multispectral satellite imagery using a vision transformer, *Nature Communications*, 15, 3801, 2024.
- 1065 Schuit, B. J., Maasackers, J. D., Bijl, P., Mahapatra, G., Van den Berg, A.-W., Pandey, S., Lorente, A., Borsdorff, T., Houweling, S., Varon, D. J., et al.: Automated detection and monitoring of methane super-emitters using satellite data, *Atmospheric Chemistry and Physics Discussions*, 2023, 1–47, 2023.
- 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.
- 1070 Tang, T., Cheng, T., Zhu, H., Ye, X., Fan, D., Li, X., and Tong, H.: Quantifying instantaneous nitrogen oxides emissions from power plants based on space observations, *Science of the Total Environment*, 938, 173 479, 2024.
- Tilstra, L. G., De Graaf, M., Trees, V. J., Litvinov, P., Dubovik, O., and Stammes, P.: A directional surface reflectance climatology determined from TROPOMI observations, *Atmospheric Measurement Techniques*, 17, 2235–2256, 2024.
- 1075 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>, 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>, accessed: 2025-06-15, 2023.
- 1080 U.S. Environmental Protection Agency, Clean Air Markets Division: CAMD's Power Sector Emissions Data Guide, <https://www.epa.gov/system/files/documents/2022-07/CAMD%27s%20Power%20Sector%20Emissions%20Data%20Guide%20-%202007182022.pdf>, office of Atmospheric Programs, Clean Air Markets Division, 2022.

- Valin, L., Russell, A., and Cohen, R. C.: Variations of OH radical in an urban plume inferred from NO₂ column measurements, *Geophysical Research Letters*, 40, 1856–1860, 2013.
- 1085 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.
- Van Geffen, J., Eskes, H., Compernelle, S., Pinardi, G., Verhoelst, T., Lambert, J.-C., Sneep, M., Ter Linden, M., Ludewig, A., Boersma, K. F., et al.: Sentinel-5P TROPOMI NO₂ retrieval: impact of version v2.2 improvements and comparisons with OMI and ground-based data, *Atmospheric Measurement Techniques*, 15, 2037–2060, 2022.
- 1090 Varon, D. J., Jacob, D. J., McKeever, J., Jervis, D., Durak, B. O., Xia, Y., and Huang, Y.: Quantifying methane point sources from fine-scale satellite observations of atmospheric methane plumes, *Atmospheric Measurement Techniques*, 11, 5673–5686, 2018.
- Veefkind, J. P., Aben, I., McMullan, K., Förster, H., De Vries, J., Otter, G., Claas, J., Eskes, H., De Haan, J., Kleipool, Q., et al.: TROPOMI on the ESA Sentinel-5 Precursor: A GMES mission for global observations of the atmospheric composition for climate, air quality and ozone layer applications, *Remote sensing of environment*, 120, 70–83, 2012.
- 1095 Verhoelst, T., Compernelle, S., Pinardi, G., Lambert, J.-C., Eskes, H. J., Eichmann, K.-U., Fjæraa, A. M., Granville, J., Niemeijer, S., Cede, A., et al.: Ground-based validation of the Copernicus Sentinel-5P TROPOMI NO₂ measurements with the NDACC ZSL-DOAS, MAX-DOAS and Pandonia global networks, *Atmospheric Measurement Techniques*, 14, 481–510, 2021.