Supplement of Atmos. Meas. Tech., 19, 6099–6124, 2026
https://doi.org/10.5194/amt-19-6099-2026-supplement

Atmospheric Measurement Techniques
Open Access
EGU

Supplement of

Global variability in the detectability of power plant NO_2 plumes from space

Ruizhe Huang and Sherrie Wang

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

Abstract. This supplement provides additional material supporting the main article: annual interference-filtering statistics (Sect. S1), per-country duplicate-filtering statistics for the $CoCO_2$ catalog (Sect. S2), model training details (Sect. S3), complementary results for the item split experiments (Sect. S4), power plant split experiments (Sect. S5), feature importance and correlation analyses (Sect. S6), and case studies of algorithm performance (Sect. S7). Section, figure, and table references not prefixed by "S" refer to the main article.

S1 Annual Interference Filtering Statistics for U.S. Top 500 Emitters

Table S1. Annual TROPOMI observation filtering for U.S. Top 500 NO_x emitters (2019–2024). Retained counts reflect the modeling subset, i.e., observations with no missing input features, a paired CAMPD record, and a target 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 171 plants.

Year	Unfiltered	Retained	Removed	Removal Rate
2019	104,672	30,966	73,706	70.4%
2020	108,786	30,838	77,948	71.7%
2021	108,053	30,934	77,119	71.4%
2022	115,294	32,808	82,486	71.5%
2023	112,427	31,534	80,893	72.0%
2024	116,990	32,633	84,357	72.1%
Total	666,222	189,713	476,509	71.5%

The filtering statistics show remarkable temporal consistency, with removal rates ranging 70–72% across all years. This stability reflects the static nature of interference zones: urban boundaries and power plant configurations change slowly compared to the annual observation cycle. The retained 189,713 observations (28.5% of the original dataset) 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.

S2 Per-Country Statistics for Duplicate Filtering in the $CoCO_2$ Catalog

Of the 16,461 candidate facilities in the $CoCO_2$ catalog, our duplicate filter (plants within 3 km sharing identical reported emissions across CO_2, NO_x, SO_x, CO, and CH_4) 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.

S3 Model Training Details

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

For the model itself, weights were initialized uniformly from the range $[-1/\sqrt{\text{in\_features}},\ 1/\sqrt{\text{in\_features}}]$. We trained the model for up to 100 epochs using the Adam optimizer with a batch size of 32 and a binary cross-entropy with logits loss function. We performed a search over seven learning rates: $1\times10^{-5}, 2\times10^{-5}, 5\times10^{-5}, 1\times10^{-4}, 2\times10^{-4}, 5\times10^{-4}, 1\times10^{-3}$, selecting the configuration with the highest AUC (Area Under the Curve) on the validation set for final evaluation on the test set. An early-stopping rule with a patience of five halted training if the validation loss did not improve for five consecutive epochs, after which the best-performing model was restored. To account for training stochasticity, we ran each experiment five times and reported the mean and standard deviation of the results.

S4 Complementary Results for Item Split Experiments

S4.1 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 S1 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 (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. AUC $= 0.83$ and global 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.

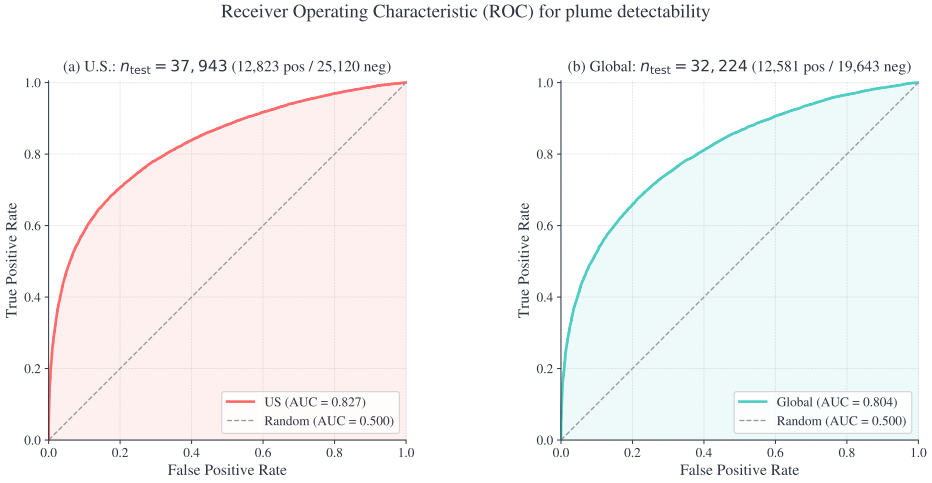


Figure S1. Receiver Operating Characteristic (ROC) curves for the U.S. (hourly NO_x) and global (annual NO_x) MLP detectability models on the "All" subset, evaluated on the held-out test split. For each region, the plotted curve is from the single best model (highest validation AUC) among five training runs; the dashed diagonal is the random-classifier baseline (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.

S4.2 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; both training and the subsequent analysis are restricted to these 171 plants.

The SHAP analysis confirms the permutation results, with NO_x emissions dominating in both the U.S. and global datasets, while adding plant-level nuance. Figure S2 maps the top feature per plant in panels (a) and (c), and the regional modal top feature on a grid 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

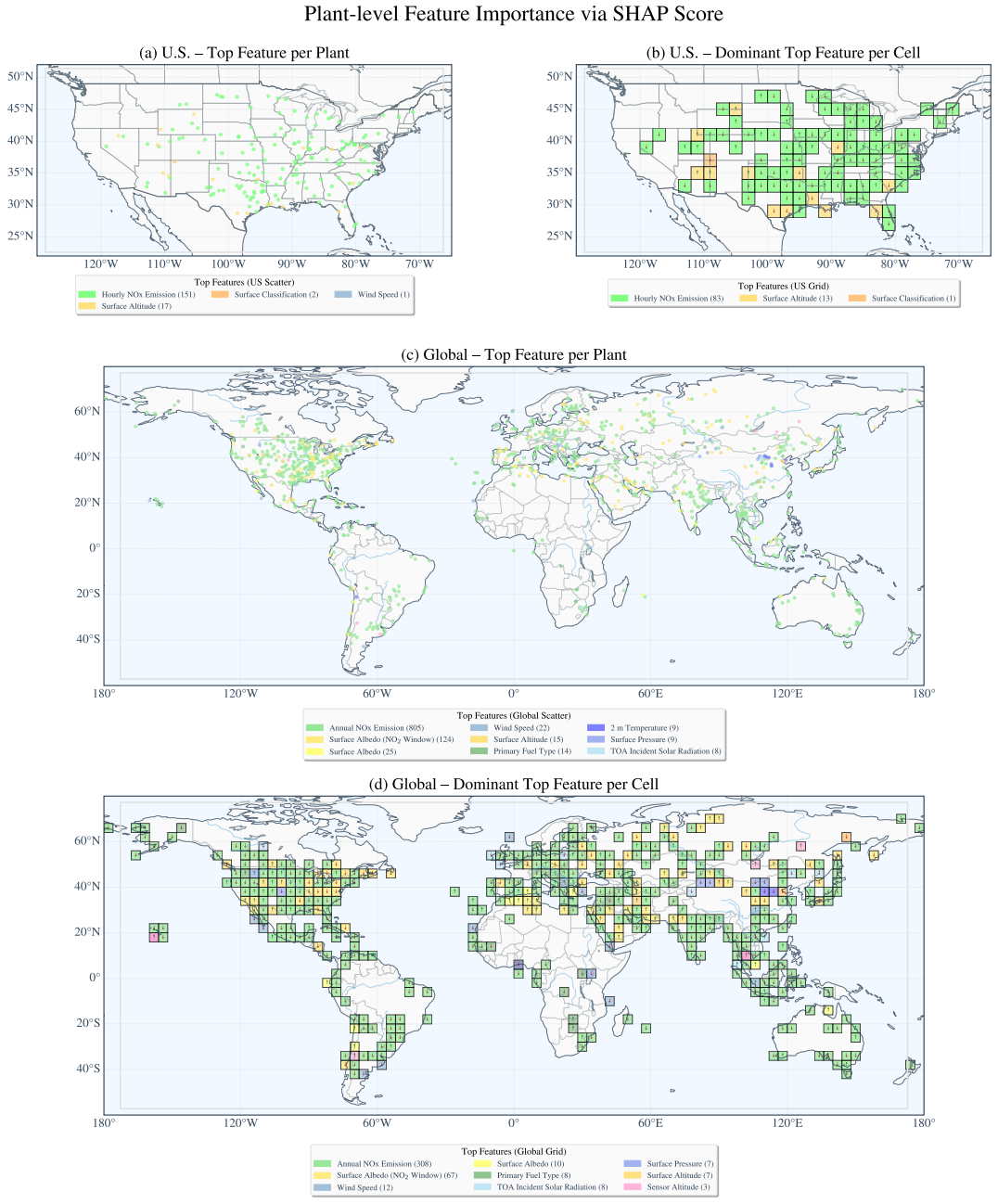


Figure S2. Plant-level feature importance via SHAP score. **(a,c)** Plant-level scatter: each point is a power plant colored by its most influential feature. **(b,d)** Grid-mode maps: within each cell (2° U.S.; 4° global), the color marks the modal top feature. Legend counts (in parentheses) report the number of plants (a, c) or grid cells (b, d) 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. SHAP values are 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.

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).

S4.3 Correlation between Per-Plant Model Performance and Key Features

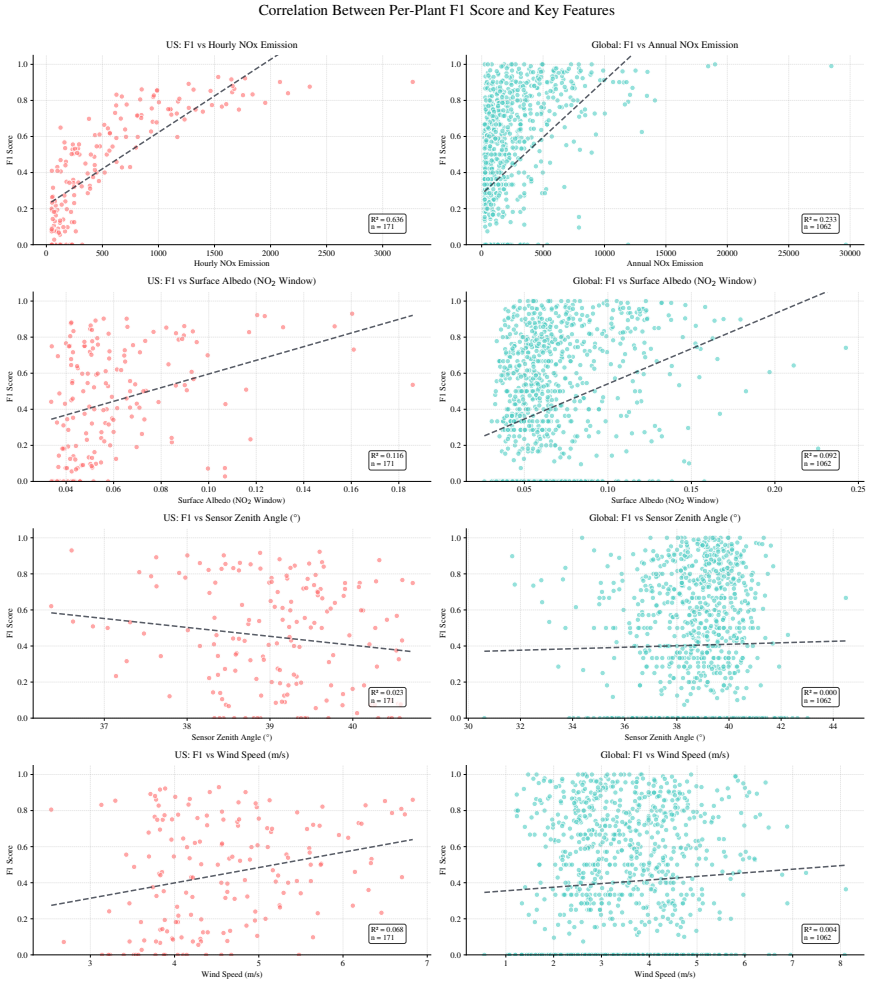


Figure S3. Correlation between F1 Score and Key Features for U.S. and Global Power Plants. Scatter plots show the relationship between the plume detection F1 score and four key features, separated for the U.S. dataset (left column, $n = 171$ plants) and a Global dataset (right column, $n = 1{,}062$ plants). F1 scores per plant are computed as in Fig. 9. We further restrict to plants with at least 3 positive observations, which is why the available power plants for global analysis drop from 1,065 to 1,062. The features (aggregated as plant-level means) are, from top to bottom: NO_x emissions, surface albedo (NO_2 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 NO_x emissions (hourly for the U.S., annual for global) in both datasets, especially in the U.S. ($R^2 = 0.636$). All other features exhibit significantly weaker correlations. 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).

S5 Power Plant Split Experiments

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

After applying the interference-zone filter, 171 power plants lie outside interference zones across every analyzed year. We retained 171 from Top 500 U.S. plants (50 from Top 100, 30 from Top 50, 13 from Top 20) and 1,065 from Top 6,000 global plants (35 from Top 100, 18 from Top 50, 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 S2 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 changes in performance for comprehensive datasets. In the U.S., accuracy on the "All" dataset is essentially preserved (0.757→0.768 for annual NO_x, 0.775→0.779 for hourly NO_x), while global accuracy declines by about 5.6 percentage points (0.746→0.690). AUC scores remain strong for the "All" models, with larger declines on the global high-emitter subsets.

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

Table S2. Multi-Layer Perceptron performance (power plant split) for U.S. and 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	Kappa
U.S.	Annual NO_x	All	0.768 ± 0.004	0.640 ± 0.021	0.446 ± 0.021	0.525 ± 0.008	0.747 ± 0.002	0.378 ± 0.004
		Top 100	0.670 ± 0.002	0.690 ± 0.009	0.716 ± 0.021	0.702 ± 0.006	0.726 ± 0.002	0.331 ± 0.006
		Top 50	0.725 ± 0.006	0.801 ± 0.005	0.793 ± 0.014	0.797 ± 0.006	0.768 ± 0.005	0.373 ± 0.009
		Top 20	0.722 ± 0.005	0.797 ± 0.017	0.827 ± 0.037	0.811 ± 0.009	0.732 ± 0.018	0.283 ± 0.037
	Hourly NO_x	All	0.779 ± 0.005	0.667 ± 0.015	0.644 ± 0.016	0.655 ± 0.003	0.812 ± 0.001	0.492 ± 0.005
		Top 100	0.742 ± 0.002	0.805 ± 0.006	0.750 ± 0.012	0.776 ± 0.004	0.811 ± 0.002	0.472 ± 0.004
		Top 50	0.777 ± 0.004	0.852 ± 0.008	0.813 ± 0.012	0.832 ± 0.004	0.832 ± 0.004	0.502 ± 0.010
		Top 20	0.761 ± 0.011	0.816 ± 0.008	0.863 ± 0.021	0.839 ± 0.009	0.783 ± 0.014	0.380 ± 0.025
Global	Annual NO_x	All	0.690 ± 0.006	0.590 ± 0.015	0.534 ± 0.021	0.560 ± 0.006	0.725 ± 0.002	0.322 ± 0.005
		Top 100	0.616 ± 0.029	0.836 ± 0.016	0.644 ± 0.042	0.727 ± 0.027	0.614 ± 0.013	0.113 ± 0.046
		Top 50	0.679 ± 0.109	0.909 ± 0.011	0.715 ± 0.143	0.793 ± 0.077	0.627 ± 0.021	0.043 ± 0.035
		Top 20	0.959 ± 0.043	0.992 ± 0.003	0.966 ± 0.045	0.978 ± 0.023	0.860 ± 0.025	0.123 ± 0.134

S6 Feature Importance and Correlation

Figure S4 shows the correlation between different variables we analyze in the U.S. Dataset and Global Dataset after filtering out the power plants in interference zones. The strongest correlations in both the U.S. and Global datasets reflect a handful of underlying physical and spatial relationships.

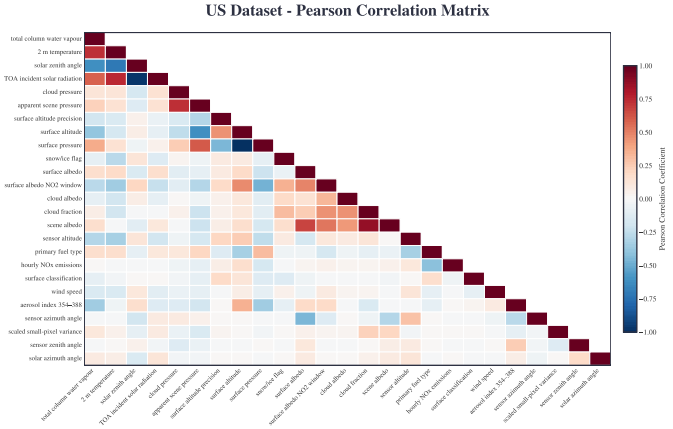
First, atmospheric variables such as surface altitude and surface pressure exhibit an almost perfect negative correlation ($\rho = -0.994$ in the U.S., $\rho = -0.986$ globally). This simply mirrors the barometric law: as you ascend, the weight of the air above you diminishes, driving pressure down.

Likewise, solar zenith angle and 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 a horizontal surface.

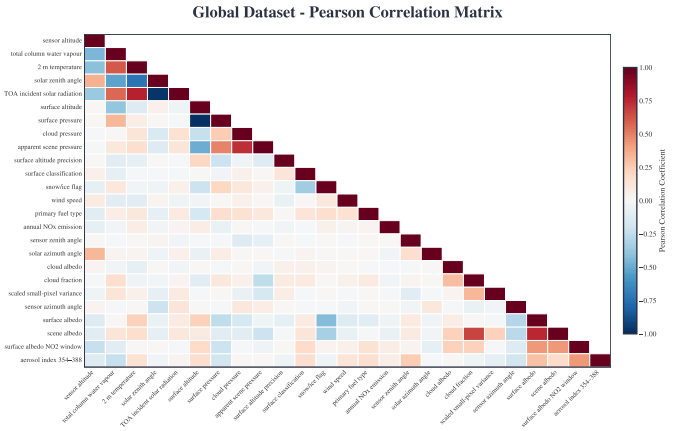
One additional U.S.-specific pattern links cloud cover to surface brightness: cloud fraction and scene albedo correlate at $\rho = 0.863$, as more cloud cover generally leads to brighter reflectance in visible wavelengths.

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

Table S3 records the permutation importance across two model configurations we displayed in Sect. 4.4. The bar chart visualization of Table S3 is Figure 10.



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



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

Figure S4. 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
Sensor Features (TROPOMI)		
sensor zenith angle	0.0295	0.0327
sensor azimuth angle	0.0029	0.0068
sensor altitude	0.0089	0.0344
scaled small-pixel variance	0.0036	0.0036
Power Plant Features		
annual NO_x emissions	—	0.1351
hourly NO_x emissions	0.2226	—
primary fuel type	0.0050	0.0493
Meteorology Features		
TROPOMI		
cloud albedo	0.0006	0.0005
cloud pressure	0.0024	0.0041
cloud fraction	0.0032	0.0090
surface pressure	0.0039	0.0276
solar zenith angle	0.0026	0.0177
solar azimuth angle	0.0009	0.0095
apparent scene pressure	0.0031	0.0110
aerosol index 354–388	0.0028	0.0086
ERA5		
100 m wind speed	0.0291	0.0226
2 m temperature	0.0034	0.0209
total column water vapour	0.0086	0.0084
TOA incident solar radiation	0.0046	0.0233
Environment Features		
TROPOMI		
surface classification	0.0095	0.0211
snow/ice flag	0.0023	0.0082
scene albedo	0.0017	0.0137
surface albedo	0.0040	0.0315
surface albedo (NO_2 window)	0.0092	0.0549
surface altitude	0.0222	0.0476
surface altitude precision	0.0056	0.0257

Table S3. Permutation importance for the U.S. "All" Hourly NO_x model and the Global "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.

S6.1 Detectability as a Function of Two Features

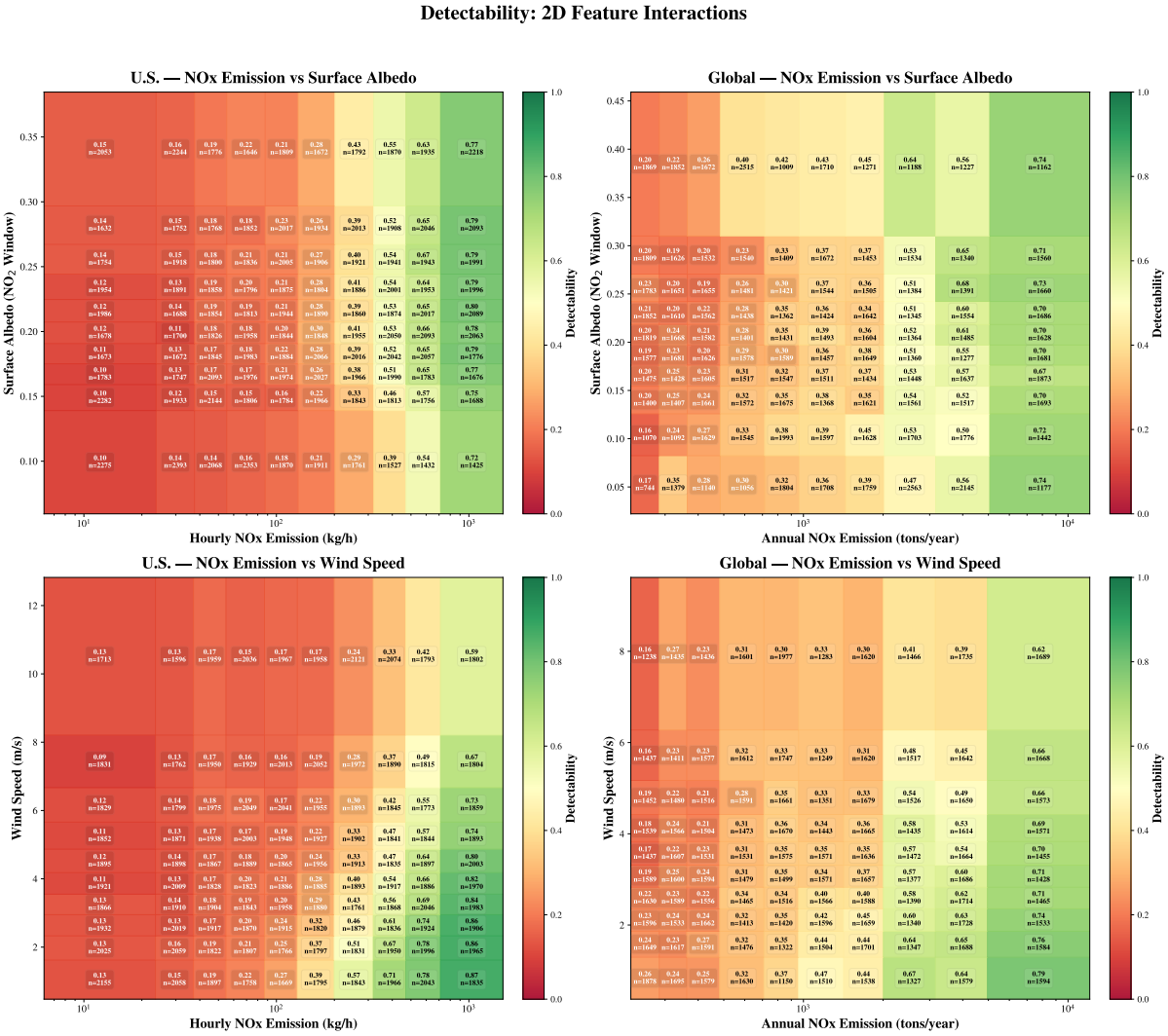


Figure S5. Two-dimensional detectability analysis for TROPOMI NO_2 plume detection from power plants. Heatmaps show empirical detectability as a function of NO_x emission rate and key environmental variables for U.S. facilities (left column; hourly emission in $\mathrm{kg\,h^{-1}}$) and global facilities (right column; annual emission in $\mathrm{t\,yr^{-1}}$). Top row: NO_x emission vs. surface albedo in the NO_2 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).

S6.2 Detectability as a function of cloud fraction

Although cloud fraction has low permutation importance in Table S3, this reflects the upstream quality flag > 0.75 filter rather than physical irrelevance. After the quality-flag filter, cloud fraction is truncated to values below ~ 0.3 (Fig. S6a), and within this surviving range detection probability still decreases monotonically with cloud fraction, from 0.41 to 0.29 in the U.S. (Fig. S6b) and from 0.45 to 0.35 globally (Fig. S6c). The cloud-fraction effect is therefore absorbed by the quality-flag filter upstream.

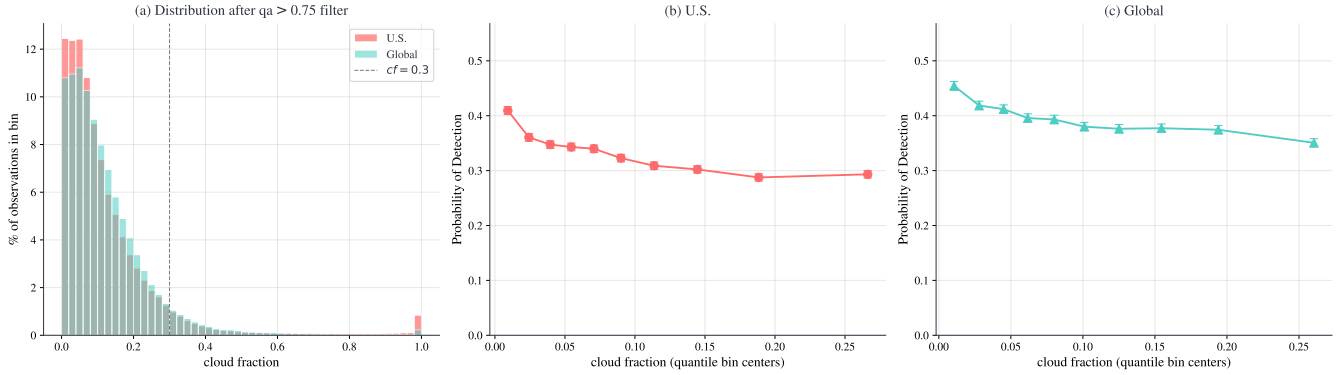


Figure S6. Detectability as a function of cloud fraction, restricted to observations passing the quality flag > 0.75 filter. **(a)** Distribution of cloud fraction in the U.S. (red) and global (teal) samples; the quality-flag filter truncates cloud fraction to values below ~ 0.3. **(b)** U.S. probability 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).

S6.3 Pixel-area quantization at the 25 km^2 flagged-area threshold

Figure S7 shows the TROPOMI pixel area as a function of sensor zenith angle (VZA). The 25 km^2 flagged-area threshold (Sect. 3.2, Step 5) is crossed at VZA $\approx 26°$: below this angle a single pixel is smaller than 25 km^2 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.

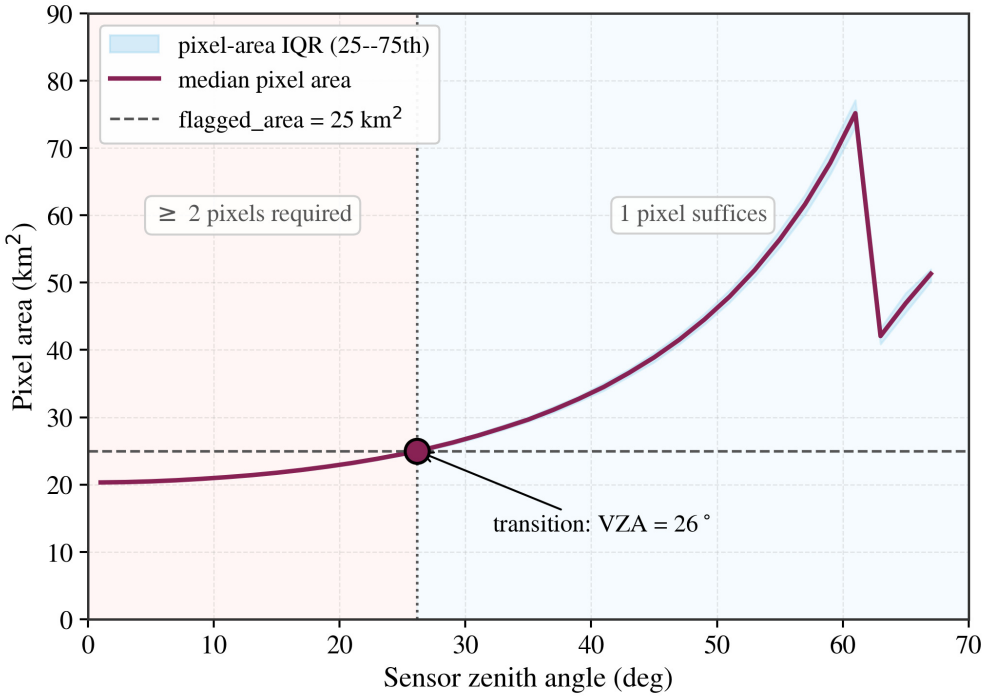


Figure S7. TROPOMI pixel area versus sensor zenith angle. The solid line is the median pixel area, with the shaded band showing the 25th–75th percentile range. The dashed horizontal line marks the 25 km^2 flagged-area threshold used in our plume-detection algorithm. Below VZA $\approx 26°$, a single pixel is smaller than 25 km^2, so detection requires at least two contiguous flagged pixels; above 26°, one pixel already exceeds the threshold.

S7 Case Studies in Algorithm Performance

To assess the performance of our Automated Plume Detection Algorithm (Sect. 3.2), we present representative examples across the four classification outcomes. A successful classification yields either a true positive (TP), where a genuine plume is correctly identified and attributed, or a true negative (TN), where the absence of a plume from the target plant is correctly noted. Misclassifications fall into two categories: a false positive (FP), where the algorithm wrongly attributes an enhancement to the target plant, and a false negative (FN), where it misses a genuine plume from the target plant.

The case studies below illustrate the underlying causes of misclassification, which include: (i) **interference**, where signals from nearby industrial facilities or urban centers are confused with the target plume, or where the interference mask itself obscures a genuine target plume; (ii) **detection threshold**, where the threshold is either too high (faint but real plumes are missed) or too low (background noise is flagged as a plume); and (iii) **regional gradients**, where natural variations in background NO_2 (e.g., land–sea contrasts or terrain-driven gradients) are misinterpreted as an emission plume.

Each case in Figures S8–S12 consists of two panels. The left panel shows the TROPOMI NO_2 field around the target plant, with the target facility (red star), wind direction (blue triangle and arrow), and 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, 2026). The header of each case lists the classification outcome, plant identifier and country, observation date, annual NO_x emissions, wind direction and speed, and the detected plume area (when applicable).

S7.1 Correct Classifications

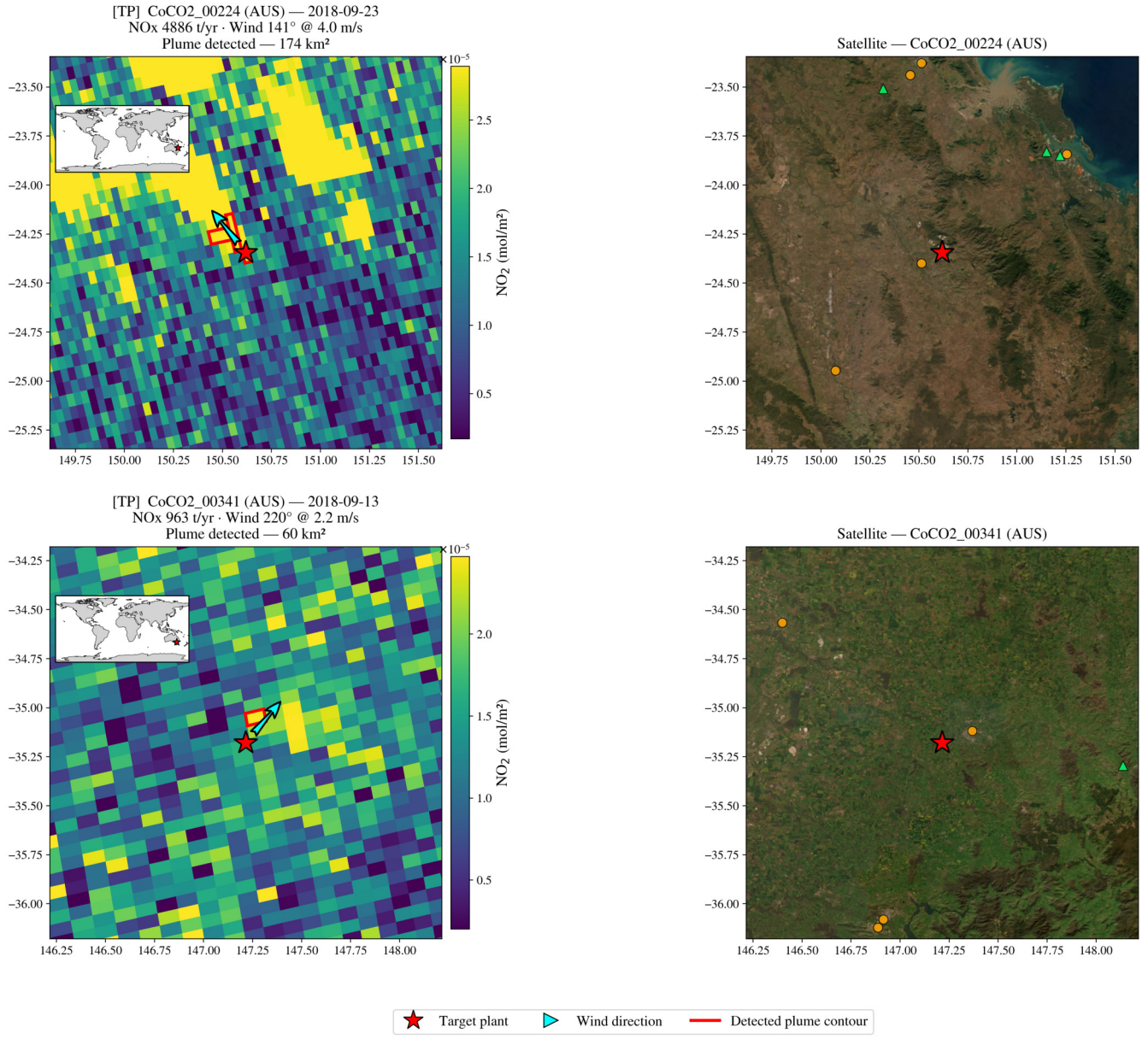


Figure S8. Two true-positive cases. The algorithm correctly identifies plumes attributable to the target plant. The top example shows a clear, well-defined plume aligned with the wind direction, while the bottom example shows a successful detection of a weaker, more diffuse plume. Basemap: ArcGIS World Imagery (Powered by Esri) (Esri, 2026) via `contextily` (Arribas-Bel and contextily contributors, 2024).

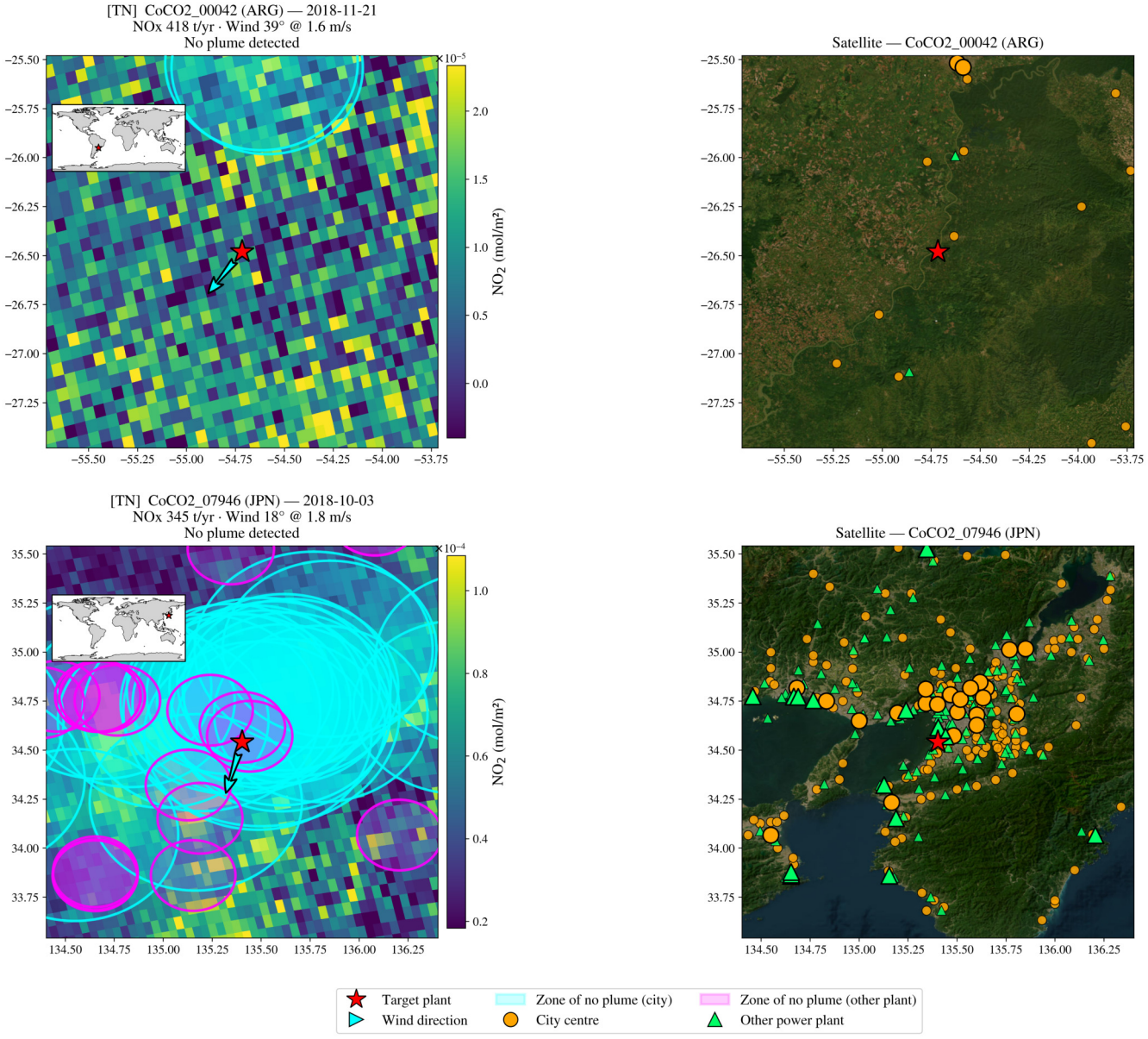


Figure S9. Two true-negative cases. The algorithm correctly returns no detection. The top example shows a quiescent scene with no discernible plume, while the bottom example shows a scene with visible NO_2 enhancements that originate from sources other than the target plant; the interference zones (cyan and pink shaded areas) correctly prevent attribution to the target. Basemap: ArcGIS World Imagery (Powered by Esri) (Esri, 2026) via `contextily` (Arribas-Bel and contextily contributors, 2024).

S7.2 Misclassifications and Their Causes

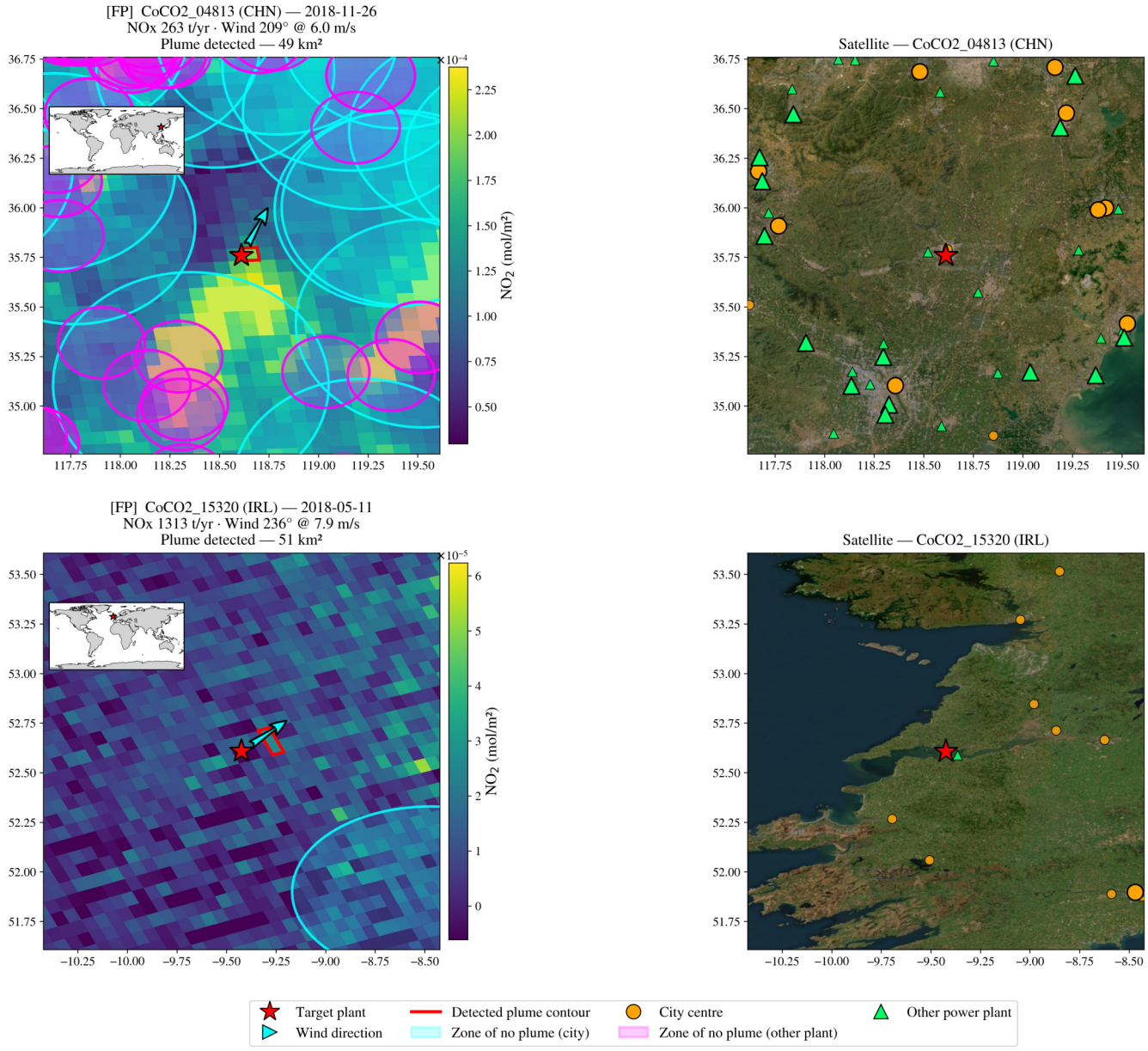


Figure S10. Two false-positive cases (part 1 of 2). The top example shows interference from a nearby urban area being mistaken for a plume from the target plant. The bottom example shows the detection threshold being set too low, so background variability is flagged as a plume. Basemap: ArcGIS World Imagery (Powered by Esri) (Esri, 2026) via `contextily` (Arribas-Bel and contextily contributors, 2024).

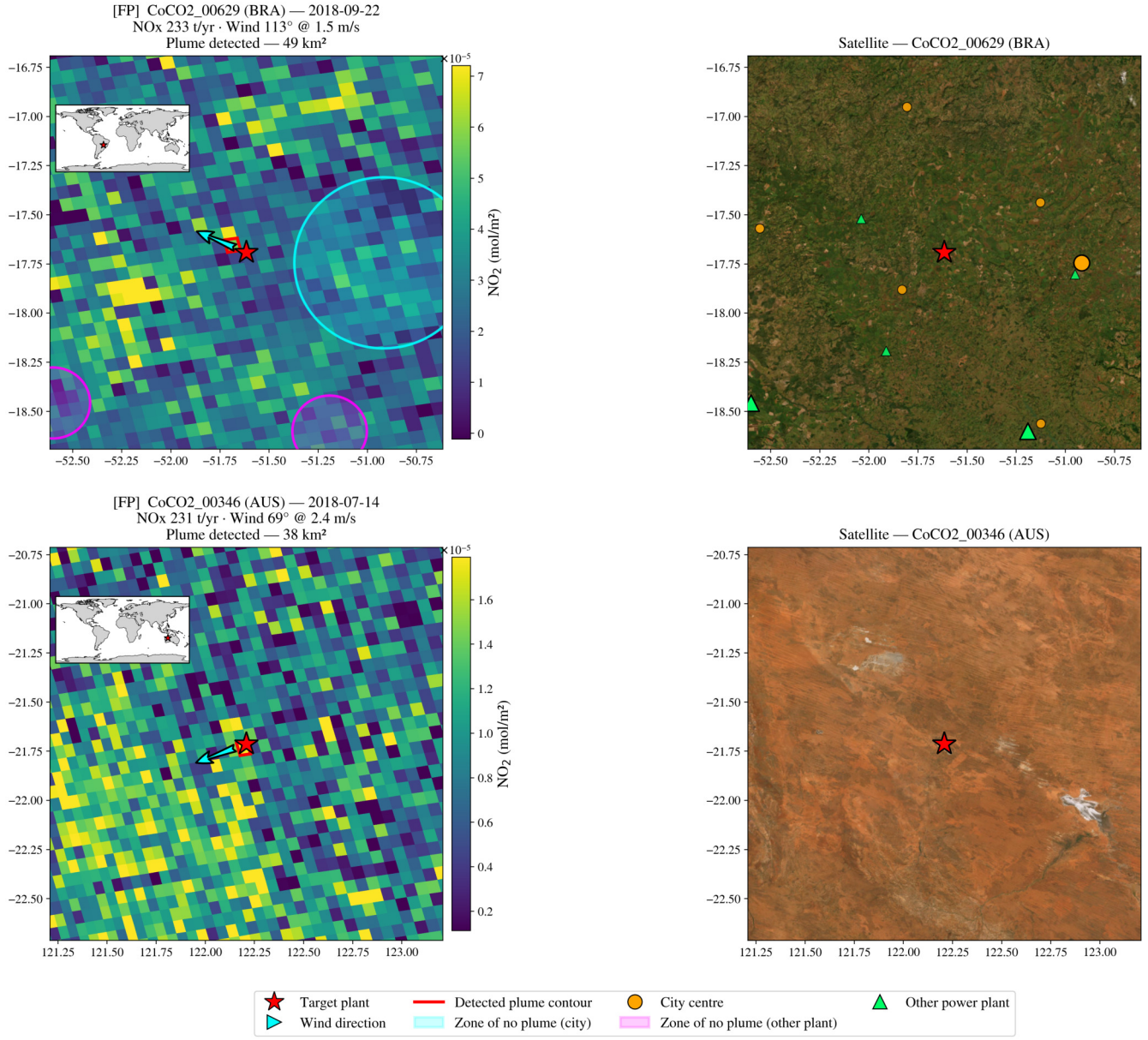


Figure S11. Two false-positive cases (part 2 of 2). Both examples show regional gradients driven by terrain being misinterpreted as a plume: the top example shows a sharp NO_2 contrast against a mountain range, and the bottom example shows a similar terrain-driven gradient producing a spurious detection in an arid region. Basemap: ArcGIS World Imagery (Powered by Esri) (Esri, 2026) via `contextily` (Arribas-Bel and contextily contributors, 2024).

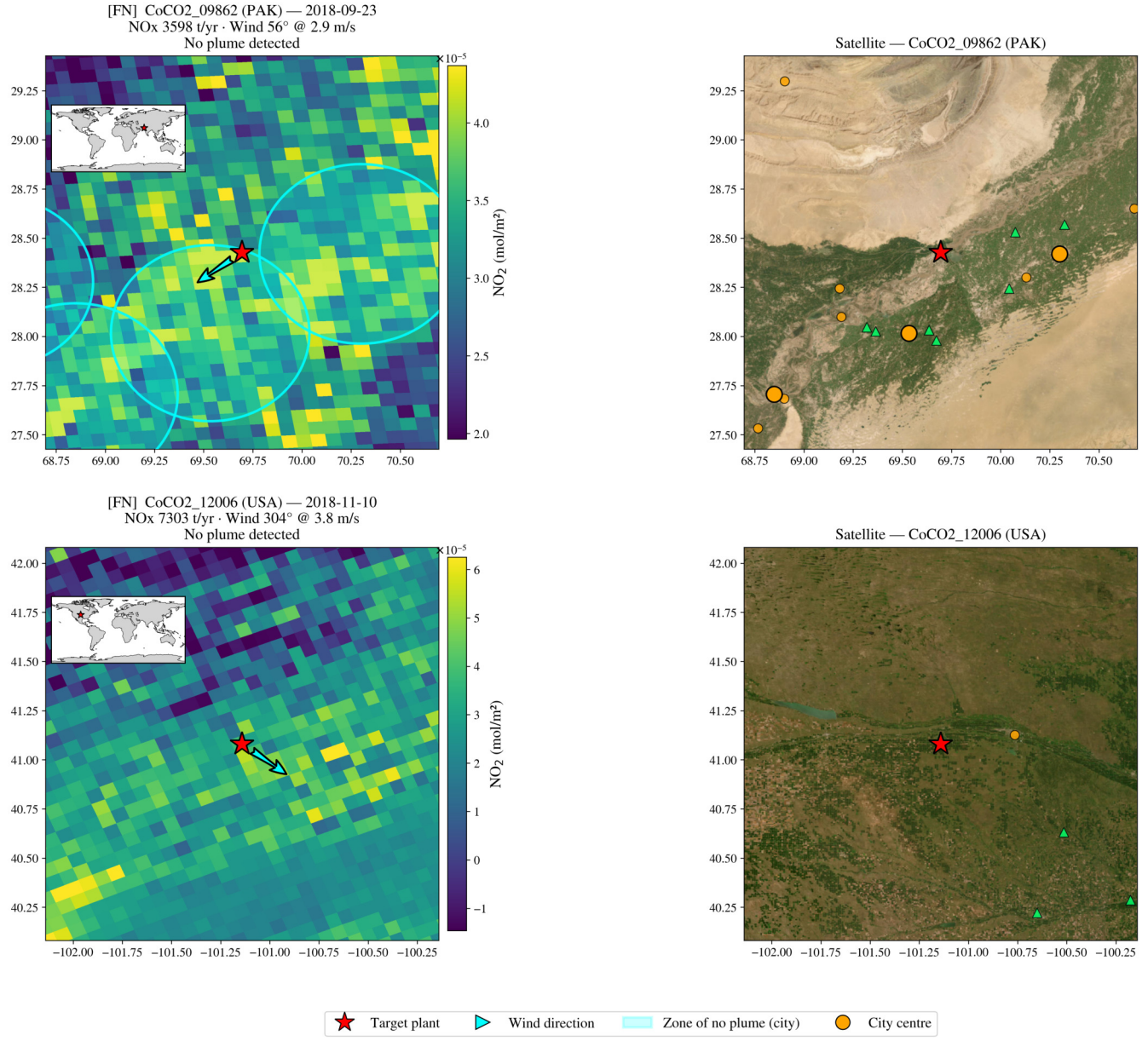


Figure S12. Two false-negative cases. The top example shows a prominent plume from the target plant that is not detected because it overlaps with a pre-defined interference mask (cyan shaded area) intended to suppress signals from nearby cities. The bottom example shows a real but low-intensity plume that falls below the detection threshold. Basemap: ArcGIS World Imagery (Powered by Esri) (Esri, 2026) via `contextily` (Arribas-Bel and contextily contributors, 2024).

References

Arribas-Bel, D. and contextily contributors: contextily: Context geo-tiles in Python, https://github.com/geopandas/contextily (last access: 10 July 2026), 2024.

Esri: World Imagery (basemap), ArcGIS Online basemap, https://www.arcgis.com/home/item.html?id=10df2279f9684e4a9f6a7f08febac2a9 (last access: 10 July 2026), 2026.