Response to Reviewers

Manuscript: *Global detectability of NO_2 plumes from power plants in single TROPOMI overpasses*

Authors: Ruizhe Huang, Sherrie Wang

Journal: *Atmospheric Measurement Techniques* (AMT)

Dear Editor and Reviewers,

We thank both reviewers for their careful reading of the manuscript, which led to clear improvements throughout. We addressed all 83 comments (31 from Reviewer 1, 52 from Reviewer 2). Below we provide point-by-point responses.

Summary of major revisions

- **Detectability is now quantitatively defined** as a binary, single-observation event (Reviewer 1 Major 3, Reviewer 2 Major 2/Major 1c/L276), with consistent terminology used throughout: *observation* is the atomic data unit (one TROPOMI overpass at a target plant plus paired variables; Section 2.2; replaces previous uses of *snapshot*), and *detection*, *detectability*, and *detection frequency* are formally defined in Section 3.
- **We replaced 10 m wind with 100 m wind** as the plume-transport variable, following Reviewer 1 Major 2/L157. We re-trained and re-evaluated the model with the new wind feature; the conclusions are unchanged but the physical justification is now correct.
- **AMF discussion in Section 5.1 and Section 5.2 has been rewritten** (Reviewer 1 L437/L481): the AMF is corrected during retrieval, so the residual sensitivity we attribute to viewing geometry is reframed as scene-dependent error in the retrieved column rather than an uncorrected AMF effect.
- **Literature review in Section 1 has been expanded** (Reviewer 1, Reviewer 2 L470/L507) with two new paragraphs covering (a) satellite top-down NO_x quantification with OMI and TROPOMI and the divergence catalogs, and (b) plume-identification methods and the theoretical scaling of downwind column enhancement. All "first" claims are now supported.
- **Section 2 has been restructured** (Reviewer 1, Reviewer 2): data sources are introduced first, the overview was moved earlier, and Section 4.1 has been relocated to Section 3 as part of the method.
- **Scope of the conclusions is tightened** to TROPOMI-specific, 7-km, isolated, mid-to-large power plants (Reviewer 2 Major 4).
- **Acronyms, references and dataset citations** have been audited and corrected (Reviewer 1 Major 1, Reviewer 2 Major 1a).

Response to Reviewer 1

1. Major 1

Reviewer comment. The study severely lacks proper references to literature and datasets. Several concrete examples are provided below. But the list below is not complete. So make sure that every dataset used in this study is properly referenced with key publications as well as working links to the datasets.

Response. We have audited every dataset cited in Section 2 to ensure each is associated with a key publication and a working link: TROPOMI L2 NO_2 (Veefkind et al., 2012; Van Geffen et al., 2020), ECMWF ERA5 (Hersbach et al., 2020), U.S. EPA CAMPD (U.S. Environmental Protection Agency, 2023), the $CoCO_2$ global power-plant inventory (Guevara et al., 2024), and the SimpleMaps World Cities Database (SimpleMaps, n.d.). Acronyms (CEMS, CAMPD, $CoCO_2$, ECMWF, etc.) are now defined consistently on first use. The literature additions are described in our replies to the comments below.

2. Major 2

Reviewer comment. The overall procedure for investigating plume detectability is comprehensible and explained well. Selected features mostly make sense, with one important exception: the 10m winds are not representing horizontal transport within the boundary layer. I see that modifying the input wind fields implies a complete re-analysis of this study. But it would be the way to go to get the best results of the presented methodology.

Response. We agree. The 10 m wind from the TROPOMI L2 product is not representative of plume transport in the boundary layer. We have replaced the TROPOMI 10 m wind with the ERA5 100 m wind field throughout the analysis. This change has been applied consistently to (i) the wind feature provided to the ML model (Table 1, Section 2.5 "Meteorology"), and (ii) the wind vector used inside the Automated Plume Detection algorithm to define the downwind cone (Section 3.2.3, Step 2). The model has been fully retrained on the new feature set, and all results, figures, and permutation/SHAP importances reported in the revised manuscript correspond to the 100 m ERA5 wind.

Quantitatively, the change has only a modest effect on the analysis. On the U.S. sample, mean wind speed increases by 34% (3.51 → 4.68 m s^{-1}) and the median per-overpass direction difference is 6.7° (75% < 15°, 90% < 33°). About 2% of binary plume labels flip (6,097 True→False and 7,453 False→True out of 666k observations), and the overall detection rate shifts only from 0.135 to 0.137. Per-plant detection rates from the two wind choices are correlated at r = 0.999.

3. Major 3

Reviewer comment. Plume detectability as defined in this study could just be put in words as 'there is an enhanced TROPOMI pixel somewhere downwind from a known power plant'. I see the motivation for this approach, and I find the study of driving features for this 'detectability' legitimate and worth to be published on AMT. However, the authors should

avoid raising wrong expectations. They should clearly state (when defining the plume detection algorithm) that this 'detectability' does not require a clear plume over several TROPOMI pixels which would be necessary for some of the emission estimation methods. They should also state (e.g. in 5.2) that this plume detectability does not provide direct information about how far and how meaningful emission estimates could be derived from these satellite measurements.

Response. We have made the scope of "detection" explicit at two locations: At the end of Section 3.2 (Automated Plume Detection), we have added: "These binary labels indicate the presence of a NO_2 enhancement attributable to the plant; they do not require a coherent multi-pixel plume of the kind used for emission quantification." At the end of Section 5.2 (Implications for emission retrieval), we have added: "We emphasize that detectability as defined here is necessary but not sufficient for emission quantification: it characterizes the upstream observational limit (whether any NO_2 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 NO_2 accumulates near the source) but is unfavorable for quantification methods that require a clearly advected plume."

4. Minor 1

Reviewer comment. Line 13: A reference to 'M et al.' is meaningless. I expect the authors refer to the paper by M. Crippa et al.

Response. Fixed. Citation corrected to Crippa et al. (2022).

5. Minor 2

Reviewer comment. Line 31: I expect that the most important difference between the four examples in Fig. 1 is the surface albedo, which is not mentioned here.

Response. Fixed. Added surface albedo (NO_2 window) to the list of drivers in the introduction. Revised: 'power plant plumes with similar NO_x emissions exhibit vastly different detectability across regions (Figure 1), driven by meteorology, surface albedo (NO_2 window), sensor geometry, and interference from nearby sources.'

6. Minor 3

Reviewer comment. Line 50: The 'NO2 Window' has not been explained yet; should be skipped, or explained.

Response. Fixed. Renamed 'Surface Albedo, NO_2 Window' to 'Surface Albedo (NO_2 Window)' on first use. Added full explanation in Section 2.5: 'Finally, 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 (Fast REtrieval Scheme for Clouds from the Oxygen A-band) 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 NO_2 fitting window (405–465 nm) for both the cloud fraction retrieval at this wavelength and the air-mass factor calculation.'

7. Minor 4

Reviewer comment. lines 52-55: Repetition.

Response. Fixed. Tightened to a single sentence: 'Together, these results clarify the observational limits of current satellite NO_2 sensors and provide empirical relationships that can inform future satellite-based emission quantification.'

8. Minor 5

Reviewer comment. How can you have exactly 400.0 kg/h for four different sources? Is this a kind of 'default' value in the database? Please check the distribution of values within your emission database.

Response. Fixed. The values are not defaults. We checked the EPA CAMPD database (503,258 non-null hourly observations, 143,292 unique values), and only 24 records (0.005%) round to 400.0 kg/h at one-decimal precision. The repeated '400.0' was a coincidence of panel selection combined with one-decimal formatting. The revised Figure 1 reports two-decimal precision (e.g. 400.08, 400.06, 399.98 kg/h), so the underlying differences between panels are visible. We have also reworded the caption to 'closely matched hourly emissions (~400 kg h-1)' instead of asserting identical values.

9. Minor 6

Reviewer comment. The legend is hard to read. In particular, '0' looks like '8' if not zoomed in. Please choose a different font without the dot inside the '0'.

Response. Fixed. The legend in the original figure used DejaVu Sans Mono, whose '0' has an interior dot. The revised figure uses Nimbus Mono PS, which has a plain '0' glyph.

10. Minor 7

Reviewer comment. The data source ('Satellite Image Source: Esri, i-cubed, USDA, USGS, AEX, GeoEye, Getmapping, Aerogrid, IGN, IGP, UPR-EGP, and the GIS User Community') is not helpful. Please clarify in the manuscript how and from where the satellite images have been derived, and give proper reference.

Response. Fixed. We have removed the long Esri attribution from all figure captions and replaced it with proper bibliographic references: the basemap is the ArcGIS World Imagery tile service (Esri, n.d.), retrieved via the contextily Python library v1.6.2 (Arribas-Bel and contextily contributors, 2024). Both references have been added to the bibliography.

11. Minor 8

Reviewer comment. Section 2: This section starts with an explanation on what has been done with the power plant data, without stating where the data is coming from first (is it

the EPA data introduced later in 2.3?). This section needs to be restructured such that the used input data is introduced, shortly described, and appropriately referenced and acknowledged first.

Response. A new opening paragraph at the start of Section 2 introduces the five input data sources with appropriate references before the subsections describe each in detail: TROPOMI Level-2 NO_2 retrievals (Veefkind et al., 2012; Van Geffen et al., 2020); U.S. EPA CAMPD records (U.S. Environmental Protection Agency, 2023); the CoCO2 global emission catalog (Guevara et al., 2024); ECMWF ERA5 reanalysis (Hersbach et al., 2020); and the SimpleMaps World Cities Database (SimpleMaps, n.d.). The Hersbach et al. (2020) ERA5 reference has been added to the bibliography.

12. Minor 9

Reviewer comment. TROPOMI NO2: references to TROPOMI (Veefkind) and the NO2 product (e.g. Geffen) are missing.

Response. Fixed. We have added both references at the point where the TROPOMI NO_2 data are introduced (Section 2.2): Veefkind et al. (2012) for the TROPOMI instrument on Sentinel-5P, and Van Geffen et al. (2020) for the TROPOMI NO_2 retrieval. Revised: '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).'

13. Minor 10

Reviewer comment. Line 80: This only holds for nadir geometry!

Response. Fixed. Original: 'Launched in October 2017, TROPOMI provides data at a spatial resolution of 7 km × 3.5 km, which improved to 5.5 km × 3.5 km in August 2019.' Revised: 'Launched in October 2017, TROPOMI provides data at a nadir spatial resolution of 7 km × 3.5 km, which improved to 5.5 km × 3.5 km in August 2019; pixel size increases toward the swath edges due to viewing geometry.'

14. Minor 11

Reviewer comment. Line 81: Note (and mention) that even if you downloaded the TROPOMI data from the NASA server, it is an ESA instrument, which also needs to be acknowledged.

Response. Fixed. Revised text in Section 2.2: 'To form the observation dataset, we used satellite data of tropospheric NO_2 vertical column densities from the TROPOMI instrument aboard the European Space Agency's (ESA) Copernicus Sentinel-5P satellite.'

15. Minor 12

Reviewer comment. Line 81: Which processor version has been used?

Response. Fixed. Added at the end of Section 2.2: 'The data come from the v2 series of the TROPOMI L2 NO_2 processor (v2.4.0–v2.8.0), spanning both the reprocessed (RPRO) and

operational offline (OFFL) streams.' The bulk of the dataset (~78%) is processor v2.4.0; the remaining files use v2.5.0–v2.8.0, which were released by KNMI between 2023 and 2024 with incremental updates within the same v2 retrieval framework.

16. Minor 13

Reviewer comment. Line 87: provide reference to qa value and 0.75 threshold.

Response. Fixed. Added Van Geffen et al. (2022) at the end of the filter description. Revised: 'a quality flag greater than 0.75, which removes data contaminated by clouds or other errors (Van Geffen et al., 2022),'

17. Minor 14

Reviewer comment. Line 144: Please specify if this equals the 'viewing zenith angle' provided in the TROPOMI L2 data. Note that viewing zenith angle also determines the size of the ground pixel, which will likely have strong impact on plume detectability!

Response. Fixed. The 'sensor zenith angle' used as a feature is the viewing zenith angle (VZA) provided in the TROPOMI L2 NO_2 product; we made this explicit in the Sensor paragraph of Section 2.5 and added a sentence noting that VZA controls the across-track ground-pixel footprint (~3.5 km at nadir to ~14 km at the swath edge) and is itself a likely driver of plume detectability. Revised passage: '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. 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.' We also replaced our previous use of 'SZA' for sensor zenith angle with 'VZA' throughout the Results section (see L436).

18. Minor 15

Reviewer comment. Line 145: How strong does the sensor altitude vary? I would expect a very low variation here that probably has no significant impact on the observation quality. Rather, I suspect that this codes in fact a latitude dependency for the NN.

Response. We agree that sensor altitude varies only modestly and that, empirically, it is well explained by latitude in our data. Sentinel-5P operates in a near-circular sun-synchronous orbit at ≈824 km. In our U.S. cohort (24°–50°N) the altitude varies over 829.0–835.3 km (≈0.8% of the mean), with Pearson r = 0.92 against latitude (r^2 = 0.85). In the global cohort the variation is somewhat larger (828.6–846.6 km, ≲ 2% of the mean) and r^2 = 0.41 against |lat|, the residual reflecting orbital geometry beyond pure latitude. The direct effect of a ≲ 2% altitude change on ground-pixel size and swath geometry is small, so the feature is retained primarily as a latitude/orbit-geometry proxy that the model can pick up alongside the explicit solar and viewing angles. We have noted this interpretation in Section 2.5: "Sensor altitude varies only modestly along Sentinel-5P's near-circular sun-synchronous orbit (≲ 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."

19. Minor 16

Reviewer comment. Line 157: Why have 10m winds been chosen? I suspect from convenience (as this is provided in the TROPOMI L2), but ERA5 data is used as well anyhow. 10m winds are already inappropriate at the source due to the high power plant stacks and the plumes' buoyancy. For describing transport to the next TROPOMI pixel ~5 km away, typical boundary layer winds are relevant, not those at the ground! I see this as a severe weakness of the study and I would expect better results for appropriate wind fields.

Response. We agree, and have addressed this issue together with Reviewer 1 comment Major 2. The TROPOMI 10 m wind has been replaced with the ERA5 100 m wind field throughout the analysis, both as the ML feature (Table 1, Section 2.5 "Meteorology") and as the wind vector in the Automated Plume Detection downwind cone (Section 3.2.3, Step 2). The model has been retrained, and all results, figures, and feature-importance values in the revised manuscript use the 100 m ERA5 wind. Please see our response to Major 2 for the full justification.

20. Minor 17

Reviewer comment. Line 161: Solar zenith angle also affects actinic flux which is key in photochemistry, affecting photolysis of NO2 directly. In addition, photochemistry affects levels of O3 (needed for conversion of NO to NO2) and OH (sink of NO2).

Response. We agree that solar zenith angle carries information about plume photochemistry beyond viewing geometry. We have updated the Radiative Transfer bullet in Section 2.5 (and renamed it "Radiative Transfer and Photochemistry") to include this: "Solar zenith angle also affects the actinic flux, which is key in photochemistry: it directly drives NO_2 photolysis, and photochemistry in turn affects the levels of O_3 (which converts NO into the NO_2 that TROPOMI observes) and OH (a sink of NO_2)."

21. Minor 18

Reviewer comment. Line 164: TOA is strongly related to solar zenith angle and basically codes seasonality.

Response. We agree that TOA incident solar radiation is strongly related to solar zenith angle and primarily reflects seasonal/latitudinal variation. We have revised the description of this feature in Section 2.5 to make this explicit: "Top-of-atmosphere (TOA) incident solar radiation 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."

22. Minor 19

Reviewer comment. Line 194: Given the number of features, I consider a sample size of 100 to be quite low for the hyper-parameter tuning.

Response. We have clarified the tuning procedure in Section 3.2. Two points are relevant: 1. Tuning set size. The tuning set actually contains 200 manually labeled observations, not 100: 100 from the global dataset (stratified across six continents and five emission quantiles, ~3–4 per cell) and 100 from the U.S. dataset (stratified across five emission quantiles, ~20 per bin). 2. The previous wording grouped these in a way that may have read as 100 only; we have rephrased to make the total 200 explicit. What is being tuned. The 25 features mentioned in the manuscript belong to the downstream MLP model (Section 3.4), which is trained on hundreds of thousands of observations (189,713 U.S. and 161,118 global, after applying the dropna criterion). The hyper-parameters tuned on the 200-observation set are the parameters of the upstream Automated Plume Detection algorithm (Section 3.2): in total ~12 parameters such as the City Masking radius, Power Plant Masking radius, wind-direction tolerance, minimum plume area, and the dual detection thresholds. These are physically motivated and the tuning set was used to qualitatively confirm and refine the chosen values rather than to perform exhaustive grid search.

23. Minor 20

Reviewer comment. Section 3.1.2: How many of the initial 500/6000 power plants are left after filtering for interferences?

Response. Fixed. This information is reported in Section 4.1. For the U.S., 171 of 500 plants were never within an interference zone over the six-year record (45.0% of NO_x emissions). Globally, 1,065 of 6,000 plants remained outside interference zones in 2018 (21.1% of emissions). We have added a forward reference at the end of Section 3.1.2: 'After this step, 171 of 500 U.S. plants and 1,065 of 6,000 global plants pass the interference filter (Section 4.1).'

24. Minor 21

Reviewer comment. Fig. 7: Since the total emissions are an obvious and direct cause for plume detectability, I would suggest to swap x and y axes in this plot.

Response. We have swapped the axes in Figure 7 so that NO_x emission rate (cause) is on the x-axis and detection count (effect) is on the y-axis, following standard convention.

25. Minor 22

Reviewer comment. Line 436: It is the first time for me to read about the sensor zenith angle - in the TROPOMI data product, there is a viewing zenith angle provided, typically abbreviated as VZA. The authors might have reasons to use the term sensor zenith angle instead. But in context of TROPOMI, the sensor zenith angle alias viewing zenith angle must not be abbreviated as SZA, as this is a standard abbreviation for the solar zenith angle.

Response. Fixed. We confirm that the sensor zenith angle in our analysis is the viewing zenith angle (VZA) provided in the TROPOMI L2 product, and we have replaced all four in-text 'SZA' abbreviations in Section 4.4 with 'VZA'. The first introduction in Section 4.4 now reads: 'By contrast, the sensor zenith angle (also known as the viewing zenith angle, VZA) shows a hump-shaped dependence…' The same equivalence is also stated where the

variable is first introduced in Section 2.5. The full term 'sensor zenith angle' is retained in the abstract, Table 1, and figure captions.

26. Minor 23

Reviewer comment. Line 437: The rise of the AMF is not really an argument here, as this is corrected for in the retrieval. In addition, pixel size increases with VZA, which should cause a decreasing detectability with VZA. Please extend the discussion here, which might be not fully conclusive; I have no good explanation for the initial increase, instead I would have expected an overall decrease with VZA.

Response. Line 437. We thank the reviewer for the correction. AMF is indeed divided out in the retrieval and pixel growth alone predicts a monotonic decrease with VZA; the original slant-path argument has been removed and the explanation revisited. Two effects can explain the rising flank. (i) Retrievals measure NO_2 slant column density along the line of sight and convert it to vertical column density by dividing by the air-mass factor (AMF); at higher VZA the longer slant path yields a larger AMF, so the same slant-column noise translates into a smaller vertical-column noise. Lower background noise raises detection probability at fixed plume signal. (ii) Our 25 km^2 flagged-area threshold introduces a pixel-quantization step. TROPOMI pixel area grows from ~20 km^2 at nadir to 25 km^2 near VZA = 26° ; detection therefore requires ≥2 contiguous flagged pixels below 26° but only one above, producing a step increase. At larger VZA, detectability falls as radiance drops, multiple scattering grows, the larger pixel size dilutes the plume signal more strongly, and retrieval uncertainty increases under oblique geometry. Section 4.4 has been rewritten accordingly, and a new appendix figure shows the pixel-area / threshold geometry.

27. Minor 24

Reviewer comment. Line 471: But the cited studies provide emission estimates, not just binary information about plume detectability.

Response. We agree that the cited studies provide emission estimates rather than only detectability information, and have revised the framing in Section 5.1 accordingly: "Previous satellite analyses of power plants have typically examined a handful of large facilities under selected conditions to estimate emissions (Goldberg et al., 2019; Beirle et al., 2019; Liu et al., 2020; Cusworth et al., 2021)."

28. Minor 25

Reviewer comment. Line 481: See Line 437.

Response. This point is addressed in our response to L437 above.

29. Minor 26

Reviewer comment. Line 515: Vertical wind shear could be extracted from ERA5.

Response. We agree that vertical wind shear can be derived from ERA5 winds on pressure levels, and we have revised the limitations paragraph in Section 5 accordingly. The text now reads: "…some variables that plausibly influence detectability were not included in our

model, including vertical wind shear (derivable from ERA5 winds at multiple pressure levels), boundary-layer mixing proxies from ERA5, and aerosol optical depth from a separate product such as the CAMS reanalysis (Inness et al., 2019).” We did not add wind shear as a feature in this revision, but note it as a direction for future work.

30. Minor 27

Reviewer comment. Fig. E1: This figure is hard to read due to the colormap (light read could be ~0 or ~1). Please use a more appropriate colormap here.

Response. We have regenerated Figure E1 with a diverging colormap centered at zero, so that values near 0 (white/neutral) are clearly distinguished from values near ±1 (saturated red/blue).

31. Minor 28

Reviewer comment. Table E1: From all listed features, I am surprised by the low number for cloud fraction. By filtering for qa>0.75, large cloud fractions are removed. But the ‘clear’ pixels are often still affected by clouds, and I would expect that detectability really depends on cloud fraction. It would thus be quite interesting to see a figure similar to Fig. 9 for cloud fraction. Please extend the discussion accordingly.

Response. Cloud fraction does affect detectability within the surviving sample. After the qa > 0.75 filter, cloud fraction is truncated to values below 0.3, yet detection probability still decreases monotonically across this range: 0.41 → 0.29 in the U.S. and 0.45 → 0.35 globally. The low importance of cloud fraction in Table A2 therefore reflects the qa filter, not physical irrelevance. Permutation importance measures a predictor’s marginal contribution within its observed support; once qa has removed high-cf cases, the residual variation is small and largely redundant with correlated cloud-related features (cloud albedo, cloud pressure, scene albedo, apparent scene pressure). The cloud-fraction effect is therefore absorbed upstream by qa and downstream by these correlated features, rather than assigned to cloud fraction in the analysis. We have added a new appendix subsection (Section F.4 “Detectability as a function of cloud fraction”) and corresponding figure showing this directly, and a one-sentence pointer in the main text at the end of Section 4.4.

Response to Reviewer 2

1. Major 1a

Reviewer comment. Abbreviations often go undefined (e.g. DDEQ, TROPOMI ect.) or used before they are defined (e.g. SHAP), also in the abstract.

Response. Fixed. We have expanded all problematic acronyms at first use: TROPOMI -> ‘TROPospheric Monitoring Instrument (TROPOMI)’ (abstract, on first use); CEMS -> ‘Continuous Emissions Monitoring Systems (CEMS)’ (abstract, on first use); OMI -> ‘Ozone Monitoring Instrument (OMI)’ (introduction); SHAP -> ‘SHapley Additive exPlanations

(SHAP)' (introduction, on first use; previously only defined in Section 3.4); DDEQ -> 'Data-Driven Emission Quantification (DDEQ) toolkit' (Section 3.2); E-PRTR -> 'European Pollutant Release and Transfer Register (E-PRTR)' (Section 2.4); LCP -> 'Large Combustion Plants (LCP)' (Section 2.4); eGRID -> 'Emissions & Generation Resource Integrated Database (eGRID)' (Section 2.4); TOA -> 'top-of-atmosphere (TOA)' (Table 1 caption / first use). Subsequent uses retain the short form throughout.

2. Major 1b

Reviewer comment. Figures and Tables need to be properly introduced and not in brackets for the first reference.

Response. All references to figures and tables in the main text (both first and subsequent references) have been converted to running-text form (e.g., "FigureX shows…" or "as shown in FigureX").

3. Major 1c

Reviewer comment. There is a general theme of inconsistency which makes for very difficult reading. The authors use the word observation, snapshot, overpass and samples, and refer to them interchangeably across figure captions and main body text (Fig 2 and sec 2.2 etc), without ever explicitly stating what these mean. These all need to be better defined and consistently used. Similarly in Section 3 the plume detection is interchangeably referred to as Automated Plume Detectability and Automated Plume Detection. It is also sometimes referred to as Automated Plume Labelling Algorithm. This confuses two main aspects of the method. The authors must be more precise in their definitions and consistent with their use throughout the manuscript.

Response. Fixed. (1) Atomic data unit. We added a definition at the end of the second paragraph of Section 2.2: 'Throughout this paper we use observation to refer to one TROPOMI overpass at a target power plant: the L2 NO_2 retrieval at the pixel closest to the plant's center, paired with the corresponding emission record.' We then use 'observation' consistently for this unit throughout the manuscript. 'Snapshot' has been replaced everywhere it referred to the same thing. 'Overpass' is reserved for the satellite passing event, and 'sample' only for standard statistical usage; previous prose uses of 'samples' for the data records (e.g. '504,165 samples') have been changed to 'observations'. (2) Plume detection algorithm. We standardized on 'Automated Plume Detection' for the algorithm that produces the binary plume label, and reserve 'plume detectability' for the predictive task. 'Automated Plume Detectability' and 'Automated Plume Labelling Algorithm' no longer appear in the manuscript (or in the box of Figure 3).

4. Major 2

Reviewer comment. How is detectability quantitatively defined? This is the key focus of the paper. What quantity is the machine learning model predicting? This fundamental definition is lacking. Suddenly in L365 detection probabilities are reported without a definition of this metric in Section 3. L368 hints that this has something to do with the number of satellite overpasses. In Section 4.4 and Figure 10 this variable appears to be

labelled as P(detect) or elsewhere P(detection). A clearer distinction needs to be made between detection probability and detectability. If these two are the same the confusion can be resolved by being more consistent.

Response. We have added a "Definitions" paragraph at the start of Section 3 that formally introduces three quantities: "Detection, a binary label produced by the Automated Plume Detection algorithm for a single TROPOMI observation: 1 if a statistically significant NO_2 enhancement is attributable to the target power plant in the downwind sector, 0 otherwise. Detectability, the probability of detection P(detect | x) given a feature vector x that characterizes the observation conditions, abbreviated P(detect) when context is clear. Detectability is estimated either empirically (as the per-plant fraction of observations with a detection) or via a machine learning model trained on the binary detection labels. Detection frequency, the absolute per-plant count of detections over an observation period (e.g., number of detections per year)."

5. Major 3a

Reviewer comment. An issue that is not discussed enough is the over/under reporting of emissions by facilities compared with what is observed in satellite data. In this study the reported emissions are used, however many studies show that there are large disagreements between these and satellite observations. Furthermore, there are also disagreements between emission inventory databases themselves, such as EDGAR and E-PRTR. The authors very briefly touch upon this at the end of Section 4.3. Given that plume detectability is derived from satellite data, this discrepancy between reported and observed (satellite) emissions must be addressed. Furthermore, NOx emission is the most important feature contributing to detectability.

Response. Our study is a detectability analysis, not an emission quantification: quantification studies (e.g., Beirle et al. 2021; Goldberg et al. 2019) derive emission rates Q from satellite observations and assess agreement with reported inventories, whereas we treat reported Q as an input feature and predict P(detection | Q). Agreement between reported and satellite-derived emissions is therefore not a premise of our methodology. Inventory uncertainty does propagate into our predicted detectability. However, this kind of measurement noise can only push a predictor's apparent importance downward, not upward (the standard regression-dilution effect: a noisily-measured predictor always looks less important than it truly is). The top SHAP ranking of reported Q is therefore a conservative lower bound, and cleaner inventories would only reinforce its dominance. We have added this discussion as a new Section 5.3 "Robustness to inventory uncertainty" in the revised manuscript.

6. Major 3b

Reviewer comment. There is no mention of the dependency on spatial resolution. I can imagine that these results would differ significantly for better spatial resolutions e.g. 1-3 km for focus mode of GOSAT-GW NO2 observations. Can the authors comment on this? If it can be shown that the key features of variability are similar across different satellite spatial resolutions, this would increase the impact of these results significantly.

Response. We acknowledge that spatial resolution influences detectability and would shift the absolute thresholds we report. However, the qualitative dependences we identify should be preserved across resolutions because they follow from instrument-agnostic mechanisms and physical process. As an internal example, the U.S. (2019–2024) and global (2018) pipelines in our study already operate at slightly different TROPOMI pixel sizes (5.5×3.5 km after August 2019 versus 7×3.5 km before) yet the two runs produce closely matching per-plant detectability profiles on the 160 plants common to both samples (Pearson r = 0.96; Appendix E) and similar trend on the top features. Finer pixels, such as the 1–3 km GOSAT-GW, would primarily shift quantitative results. Because more plants reach a detectable signal, the positive class is better populated, which would raise the achievable model performance above the F1 and AUC reported here. The relative ranking of features should nonetheless remain (NO_x emission rate dominant, with surface, geometry, and meteorology as second-order drivers), because these dependencies follow from instrument-agnostic physics. Conducting an analogous analysis on other sensors is beyond the scope of the present study and is a natural direction for follow-up work. We have added a corresponding note in the limitations paragraph of the revised manuscript, citing the GOSAT-GW mission.

7. Major 3c

Reviewer comment. For the global dataset, the use of a single emission value for almost an entire year of datapoints seems questionable. On the other hand the hourly data available from the US dataset would be more robust. Is there sufficient stability in the hourly emissions timeseries of a single power plant to justify using one value for the global data? Are there overlapping power plants in both datasets for which the predicted detectability values can be compared? Could this be a reason for the systematically higher detectability in the global dataset compared to the US only one?

Response. Hourly emissions are generally preferable to annual values where they exist, and we use them for the U.S. analysis (CEMS hourly NO_x). For the ~6,000 plants in our global sample we are not aware of a comparable per-plant hourly record, so we adopt the CoCO2 inventory (Guevara et al., 2024). Compared with the hourly rate, the annual mean is a less faithful proxy at the level of an individual observation. Such noise on a predictor tends to attenuate rather than amplify its apparent contribution. The SHAP ranking of Q in our model can therefore be read as a conservative estimate of its true importance, and access to global hourly data would, if anything, sharpen rather than weaken this signal. Following the reviewer's suggestion, we directly compared the two pipelines on physically identical plants. Matching within <10 m yields 160 plants present in both runs (1,273 vs. 137 observations/plant on the U.S. and global sides). Per-plant detectability is in close agreement: Pearson r = 0.96; 45% of plants match within $|\Delta| < 0.01$. The small residual offset (mean +0.036, U.S. mean 0.163 vs. global 0.199) may reflect the temporal gap between the two samples, since 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. We have added this overlap analysis as a new figure in Appendix E ("Cross-pipeline consistency on overlapping plants") of the revised manuscript.

8. Major 4

Reviewer comment. In light if the above comments on the discussion, the scope of the conclusions and abstract should be somewhat reduced. It should be specified that these results are for a spatial resolution of around 7 km (S5P pixels) and that the trained model is sensor specific so its output could only be used when looking at TROPOMI data. Furthermore, when presenting a 98 % plume detection accuracy, it should be immediately added that this is for isolated plumes that are 20 km away from other power plants and up to 90 km away from cities. Finally, the caveat that only 21.1 % and 45.0 % of total emissions are accounted for in this analysis, for the global and U.S datasets respectively, must be stated in both the abstract and conclusion.

Response. We have revised both the abstract and conclusion to add the requested caveats, with one clarification: the 98% plume-detection accuracy was measured on 400 manually labelled observations randomly drawn from the full test set (Section 3.2), which spans both isolated plants and plants near other emission sources. We therefore did not attach the isolation criteria directly to the 98% figure. The revised abstract now reads: "We present the first global, data-driven analysis of power plant NO_2 plume visibility from space. Using 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 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 (the probability of detection given the observation conditions) from environmental, meteorological, sensor, and power-plant variables sampled at the single TROPOMI pixel over each plant ($F1 > 0.66$, $AUC > 0.8$). […]" And the conclusion's first paragraph: "In this work, we systematically mapped power plant NO_2 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. […] Because the model is trained on TROPOMI features, its predictions apply to TROPOMI-like observations rather than to satellite NO_2 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. […]"

9. Minor 1

Reviewer comment. The presentation of the datasets in Section 2 is very confused. Is the first half of Section 2.1 describing the same dataset as Section 2.3? Further, the second paragraph of 2.1 is a repeat of Section 2.4. The level of detail of filtering out power plants is not needed. Please remove Section 2.1 altogether.

Response. We have removed Section 2.1 in its entirety, since its content was indeed redundant with Section 2.3 (U.S. cohort) and Section 2.4 (global cohort). Specifically: The U.S. selection (top 500 plants, 2019–2024), fuel-mix breakdown, and coal-to-natural-gas

transition statistics have been integrated into Section 2.3 (Hourly NO_x Emissions in the U.S.). The global selection (top 6,000 plants, 2018), emission range, and fuel-mix breakdown have been integrated into Section 2.4 (Annual NO_x Emissions for Global Power Plants). The duplicate-filtering description has been compressed to a brief two-sentence summary in Section 2.4 (where it logically belongs, since the duplicate problem is specific to the global $CoCO_2$ catalog), with the per-country breakdown and a brief discussion of the underlying catalog-integration causes moved to a new appendix (Appendix B, "Per-Country Statistics for Duplicate Filtering in the $CoCO_2$ Catalog"). Figure 2 (the dataset overview figure showing observation distribution and emission histograms for both U.S. and global power plants) has been retained at the start of Section 2 as an overview figure. We have also revised the framing of the duplicate filter from "data quality issues" to "catalog-integration challenges", to clarify that duplicate entries are not necessarily indicative of errors in the underlying emission estimates.

10. Minor 2

Reviewer comment. The ordering of the subsections in Section 2 is a bit unnatural. It begins with very detailed information about datasets, while the overarching information about the datasets and what they are used for is only given at the end in 2.5. I suggest making the contents of 2.5 the main part of Section 2, and then go on to present each individual dataset in more detail.

Response. To address the ordering, we added an introductory paragraph at the start of Section 2 that lists the five input data sources and briefly notes what each is used for, so that the overview comes before the per-dataset subsections: "Our analysis combines five input data sources: Sentinel-5P TROPOMI Level-2 NO_2 retrievals, 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 for hourly NO_x emissions at 500 U.S. power plants; the $CoCO_2$ global power-plant emission catalog for annual NO_x emissions at 6,000 plants worldwide; ECMWF ERA5 reanalysis for meteorological variables; and the SimpleMaps World Cities Database 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." We kept Section 2.5 (Features Predictive of Plume Detectability) as a separate subsection because it describes how variables drawn from these sources are assembled into the feature set used by the model, rather than introducing a data source. We hope this addresses the reviewer's concern about ordering.

11. Minor 3

Reviewer comment. The contents of Section 4.1 is less a result, rather more a by-product of the method. Since it is not the focus of the paper which addresses plume detectability, I suggest changing this to be a more detailed, quantitative part of Section 3, i.e. a subsection dedicated to the training sample, unless the authors can give a good connection between this point and it's impact on plume detectability.

Response. We agree. We have moved the contents of Section 4.1 to a new Section 3 subsection "Training Sample Construction" (Section 3.3), and replaced Section 4.1 with a

brief paragraph ("Scope of the analysis") summarizing the cohort sizes (171 U.S. plants, 1,065 global plants; 45.0% and 21.1% of original NO_x emissions retained) and pointing readers to Section 3.3 for the full details.

12. Minor 4

Reviewer comment. Sporadic use of bold font throughout - particularly it appears in Section 3 and Section F. Bold font should be removed.

Response. Fixed. We removed all of the bolded font in the text except the item tag and paragraph head.

13. Minor 5

Reviewer comment. There is a disconnect between the conclusions and the rest of the main body text. For example, aerosol optical depth is mentioned in the conclusions but never anywhere else in the text. The authors must make an effort to harmonise the conclusion with the rest of the paper.

Response. The other terms mentioned in Section 6 (vertical wind shear, turbulence) are framed as future-work directions rather than current-paper limitations, and we have rephrased Section 6 to make this distinction explicit. Future-work items by their nature extend beyond the present feature set and are not introduced earlier in the manuscript.

14. Minor 6

Reviewer comment. Whilst they are visually helpful and insightful into the plume detection, the number of figures in Section F is too many (more than the whole main text) - this can be reduced by consolidating panels into less figures or by reducing the number of examples to, say, 6. Also all symbols need to be made bigger. I cannot see the wind direction icon anywhere.

Response. We have condensed and redesigned the appendix: 1. The 11 separate figures have been consolidated into 5 figures organized by classification outcome: 1 figure for true positives (2 cases), 1 figure for true negatives (2 cases), 2 figures for false positives (2 cases each, split into "part 1" and "part 2" to avoid an over-tall single figure), and 1 figure for false negatives (2 cases). The total number of cases shown is unchanged (10 examples), but they now occupy 5 figures instead of 11. 2. Each case is now displayed using two panels (TROPOMI NO_2 field + satellite imagery) rather than the previous six-panel layout, with all symbols (target plant, wind direction, plume contour, interference zones, nearby cities and other power plants) enlarged. 3. The wind direction is now shown as a clearly visible blue triangle with an arrow on the NO_2 panel of every case, and is explicitly listed in the legend of each figure. 4. The section text has also been streamlined: the previous subsubsection-level breakdown (True Positives / True Negatives / Failure Due to Interference / Failure Due to Incorrect Detection Threshold / Failure Due to Regional Gradients) has been consolidated into a single introduction that lists the misclassification causes once, with each figure caption indicating the relevant cause for the cases shown.

15. Minor 7

Reviewer comment. Formally this study deals more with enhancements rather than plumes, since there is no constraint on plume structure, such as a certain number of enhanced pixels adjacent to each other etc. I think that, given the fact that there are probably mostly real plumes in their dataset, this is ok, however it should be more clearly and explicitly stated that environmental, meteorological, and observational variables are extracted from a single pixel. This should be included in both the abstract and conclusion.

Response. Fixed. Abstract revision: 'from environmental, meteorological, and observational variables' -> 'from environmental, meteorological, and observational variables sampled at the single TROPOMI pixel over each plant'. Conclusion revision: 'Despite its scope, this analysis has several limitations.' -> '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 characterises NO_2 enhancements at that pixel rather than plume-wide structure.'

16. Minor 8

Reviewer comment. The authors make repeated claims that they provide the 'first' study to address either quantifying plume visibility (L470) or demonstrate the factors that contribute to this (L488, L505) without sufficient reference to the literature. I think a clearer presentation of the science question in the introduction (Section 1), along with a more thorough discussion of the current literature would go a long way to support this claim. From L30 the introduction gets a bit lost. Until the end (L55) there are no more literature references and the text becomes about method, results and datasets.

Response. We have rewritten the introduction (Section 1) and scoped the "first" claim narrowly. Two new paragraphs place our work alongside the existing satellite NO_x quantification literature, the complementary plume-identification line of work, and the theoretical expectations on which our detectability model rests. Against this expanded background, our claim is now: to our knowledge, no prior study has combined these ingredients into a supervised, CEMS-validated, globally applied model that predicts plume detection probability across thousands of power plants. The framing makes clear that our contribution complements the divergence catalog of Beirle et al. (2023).

17. Further minor 1

Reviewer comment. L13: Typo in reference.

Response. Fixed. Citation typo corrected (see RC1-L13: Crippa et al. 2022).

18. Further minor 2

Reviewer comment. Figure 1: an additional piece of relevant information would be the date and time of the observation. Also there is an underlying dashed grey grid plotted which has a different projection than the TROPOMI pixels. This should be removed.

Response. Fixed. The acquisition date and time (UTC) of each TROPOMI overpass are now shown in the annotation box on each panel, and the dashed lat/lon grid has been removed from both the NO_2 and satellite panels.

19. Further minor 3

Reviewer comment. L75: I do not understand the supposed link between data quality and the applied filtering. Does the presence of duplicates really imply that the estimation of emissions values are worse?

Response. We have rephrased the passage so that the filter is described as flagging duplicate entries rather than implying anything about emission accuracy: "Together, China and India accounted for 66% of all removed plants (2,194 out of 3,320). These removals reflect duplicate entries, where the same plant is recorded multiple times in the catalog. The United States had only 1 plant removed out of 2,850."

20. Further minor 4

Reviewer comment. L79: Reference Veefkind et al. (2012)

Response. Fixed. Added Veefkind 2012 citation. Revised: 'To form the observation dataset, we used satellite data of tropospheric NO_2 vertical column densities from the TROPOMI instrument aboard the European Space Agency's (ESA) Copernicus Sentinel-5P satellite (Veefkind et al., 2012).'

21. Further minor 5

Reviewer comment. L90: what does observation here refer to? Pixels? or orbits or overpasses?

Response. Fixed. Defined explicitly in Section 2.2: 'Throughout this paper we use observation to refer to one TROPOMI overpass at a target power plant: the L2 NO_2 retrieval at the pixel closest to the plant's center, paired with the corresponding emission record.' (See also response to Minor 1c.)

22. Further minor 6

Reviewer comment. L92: what is a 'snapshot'?

Response. Fixed. We changed 'snapshot' into 'observation' throughout the manuscript. (See also response to Minor 1c.)

23. Further minor 7

Reviewer comment. L110: Pairing of TROPOMI data with emissions is a step too far for a datasets section. This should be moved to Section 3. Similarly for L130.

Response. Fixed. We moved these two parts (L110 and L130) into the method section under a new subsection 'Pairing TROPOMI Observations with Emissions'.

24. Further minor 8

Reviewer comment. L119: specify that the total global power plant emissions are the reported emissions.

Response. Fixed. Original: 'This comprehensive inventory covers at least 95.9% of total global power plant emissions…' Revised: 'This inventory covers at least 95.9% of total reported global power plant emissions…'

25. Further minor 9

Reviewer comment. L131: definition of incomplete data needs clarified. It needs to be more clear which data were used and which were removed and the reasons for doing so.

Response. Sections 2.2 and 3.1 have been rewritten to make explicit which observations are used at each analysis stage: Dataset characterization and detection statistics (Sections 2.2 and 4) use the full QA-passed set: 666,222 (U.S.) / 875,686 (global) observations. Model training and feature-importance analyses (Sections 3.3 and 3.4) use a restricted subset (189,713 U.S.; 161,118 global) consisting of observations with (i) no missing (NaN) values in any input feature - TROPOMI retrieval variables can themselves contain NaNs - and (ii) a paired emission record (hourly CAMPD for the U.S.; annual CoCO2 for the global analysis). The previous "removed a small number with incomplete data" has been replaced with this explicit definition.

26. Further minor 10

Reviewer comment. Table 1, footnote c: The reader is introduced to a lot of concepts that they have not yet been presented, or are not at all (ROCINN occurs nowhere else in the manuscript). This is confusing and should be mentioned in the text or at least be referenced to the section where they are all discussed.

Response. Fixed. We rewrote footnote (c). The original text introduced two algorithm names (CRB and ROCINN) without context, and the ROCINN reference was inaccurate for our dataset: in our TROPOMI L2 NO_2 v2.4–v2.8 files, the cloud_selection_flag selects FRESCO for ~99.7% of pixels and ROCINN for none. The revised footnote names the actual retrieval, defines the FRESCO acronym in place, and keeps the CRB approximation as a one-line note. Original: 'Input variables for Clouds-as-Reflecting-Boundaries (CRB) model, a simplified Lambertian-reflector approach used by the ROCINN algorithm to retrieve cloud properties.' Revised: 'Cloud-product variables provided by the FRESCO (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) approximation.' The fields involved (cloud albedo, cloud pressure, cloud fraction) are also now described in Section 2.5 alongside the two-wavelength surface albedo discussion, so the reader encounters FRESCO and CRB in the main text before reaching Table 1.

27. Further minor 11

Reviewer comment. L169: state that the albedo is at different wavelengths.

Response. Fixed. Original: 'Finally, multiple albedo measurements quantify surface reflectivity, which controls the signal strength for detecting a plume against different backgrounds.' Revised: 'Finally, multiple albedo measurements at different wavelengths quantify surface reflectivity: surface albedo and scene albedo at 758 nm, both provided by FRESCO (TROPOMI's standard cloud retrieval, which uses the O2 A-band to estimate cloud fraction and pressure), and surface albedo (NO_2 window) at 440 nm, used in the NO_2-window cloud fraction retrieval and the air-mass factor calculation. These control the signal strength for detecting a plume against different backgrounds.'

28. Further minor 12

Reviewer comment. L188: There are two main parts to Section 3, the plume detection and the training of a model to predict predictability. As a reader it is easy to confuse these two things - change plume-detectability-with-attribution to plume-detection-with-attribution. Likewise the text in the box in Figure 3 should be changed accordingly (Automated Plume Detectability Labeling to Automated Plume Detection).

Response. Fixed. We use 'Automated Plume Detection' throughout for the algorithm and reserve 'plume detectability' for the predictive task. The following replacements were applied: plume-detectability-with-attribution -> plume-detection-with-attribution; Automated Plume Labeling Algorithm -> Automated Plume Detection Algorithm; Automated Plume Detectability Algorithm -> Automated Plume Detection Algorithm; plume-labeling algorithm -> plume-detection algorithm; 'the labeling algorithm generates…' -> 'the detection algorithm generates…'; 'Plume Labeling Algorithm via Manual Labeling' -> 'Plume Detection Algorithm via Manual Labeling'; 'automated plume-labeling and machine-learning framework' -> 'automated plume-detection and machine-learning framework'. The Figure 3 box has been updated to 'Automated Plume Detection'.

29. Further minor 13

Reviewer comment. L195: Give the definition of the quantiles (every 20 % ?). Please define them here, instead of in Section 3.1.7.

Response. The five emission-level quantiles are now defined where they first appear (Section 3.1.1, Hyper-parameter Tuning): "…balanced across six continents and five emission-level quantiles (0–20%, 20–40%, …, 80–100%) by selecting roughly 3–4 observations per stratum…" The previously redundant in-line definition in Section 3.1.7 has been replaced with a back-reference.

30. Further minor 14

Reviewer comment. L198: Which specific model hyper-parameters were tuned?

Response. The tuned hyper-parameters are now enumerated at the beginning of the Hyper-parameter Tuning paragraph (Section 3.1.1): Step 1: city population threshold, plant emission-ratio threshold, City Masking radius, Power Plant Masking radius Step 2: wind-direction tolerance, maximum search distance, close-range distance, minimum plume area Step 4: upwind sector angle, background annulus Step 5: Statistical Significance (2σ), Absolute Minimum thresholds Their final values are reported in the corresponding step

descriptions. For the ML model, the tuned hyperparameter is the learning rate (selected from a 7-value grid by maximising validation AUC).

31. Further minor 15

Reviewer comment. Section 3.1: Can you mention here that the parameters used as input to the model training and prediction are extracted from a single TROPOMI pixel, and not across the entire plume?

Response. Fixed. We added '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.' after 'For each analysis, we pair every TROPOMI overpass over a selected power plant with the corresponding emission record to form the input samples used by all subsequent steps.'

32. Further minor 16

Reviewer comment. Figure 4: Please enlarge the blue arrow.

Response. We've enlarged the blue arrow in Figure 4.

33. Further minor 17

Reviewer comment. Section 3.1.3: Please give the distances and areas also in terms of number of TROPOMI pixels.

Response. We have considered adding pixel-count equivalents alongside each distance, but ultimately retained the km-only formulation for two reasons. First, TROPOMI's across-track ground footprint varies substantially with viewing zenith angle (from ~3.5 km along-track and 5.5–7 km across-track at nadir to up to ~14 km across-track at the swath edge; see Section 2.2), so a single fixed pixel-count would be misleading for any given off-nadir observation. Second, the algorithm's geometric operations (masking, downwind cones, annular background regions) are inherently defined in physical distance rather than pixel space, and we found that mixing the two units made the description harder to read. We have kept the nadir pixel size in Section 2.2 as the reference; readers can use it to convert any distance to an approximate nadir pixel count (e.g., 20 km along-track ≈ 6 nadir pixels).

34. Further minor 18

Reviewer comment. L238: Please elaborate on this special consideration within 5 km - why and how this is done.

Response. The sentence has been rewritten to make the rule explicit. Within 5 km of the plant, the ±25° wind-direction tolerance is relaxed to include any pixel that contains the plant location, since at this close range the plant can sit inside a pixel whose center is offset from the wind direction.

35. Further minor 19

Reviewer comment. L255: Can a rough value of how many scenes are filtered out for each criteria be quoted?

Response. Yes. Of the 666,222 (plant, overpass) snapshots entering the detection algorithm, the per-criterion rejection breakdown is:

- 47.3% (315,148): Step 3, the Plume Available Zone is empty, i.e., the interference mask (Step 1) and the downwind cone (Step 2) jointly leave no candidate pixels.
- 27.9% (185,649): Step 5, pixels in the Plume Available Zone fail the local 2σ statistical threshold only.
- 8.1% (53,635): Step 5, pixels fail both the 2σ statistical and 5×10^{-6} mol m^{-2} absolute thresholds.
- 3.1% (20,746): contiguous plume area below the 25 km^2 minimum (Step 2 criterion, applied after Step 5 thresholds).
- <0.1% (251): Step 5, pixels fail the absolute floor only.
- 13.6% (90,793): pass all criteria, positive detection.

Two takeaways: (i) the upstream geometric filters (Steps 1–2, combined as Step 3) eliminate nearly half the candidate scenes before any column threshold is applied; (ii) within the Step 5 dual-threshold test, the local 2σ statistical criterion is the dominant filter, while the absolute floor of 5×10^{-6} mol m^{-2} adds essentially no rejections by itself, serving as a safeguard against noise-driven false positives in low-background scenes.

This per-step audit is provided for the reviewer's reference; we have not added it to the manuscript, as we believe it would be too granular for the Methods section.

36. Further minor 20

Reviewer comment. Figure 5: Confusion matrices are typically visualised in a grid, such as Figure 7, which I think would be better. What does true and false correspond to, plume or no plume? Is no plume detected in 5b because the centroid lies within the no-plume zone or because either of the statistical significance or absolute minimum conditions were not fulfilled, or both?

Response. We have revised Figure 5 as follows: 1. Panels (c) and (d) now use colored confusion matrices (green for correct cells, red for misclassified cells), with row and column totals retained. 2. The caption clarifies that "True" denotes the presence of a plume attributable to the target plant and "False" denotes its absence, that "Predicted" is the algorithm's binary output, and that "Actual" is the manual human label. 3. The caption also notes that in panel (b) the target plant's location lies inside an interference-zone mask (no-plume zone) defined in Step1, which is why no plume is returned. Outside the mask, a detection would additionally require the Step5 dual-threshold conditions (statistical significance and absolute minimum) and the Step 2 minimum plume-area condition to be met. We have also re-validated the algorithm on the updated stratified samples: the U.S. validation gives 98.0% accuracy (precision 93.5%, recall 93.5%), and the global validation gives 98.0% accuracy (precision 88.0%, recall 95.7%).

37. Further minor 21

Reviewer comment. L276: My interpretation of plume detectability is that the integer number of detected plumes is predicted and then divided by 100. In doing so the numbers in L365 are achieved (6.02 corresponds to 602 detected plumes). Is this correct?? If the definition of plume detectability is 'plume or no plume', how is this any different from plume detection in Section 3.1 A proper, clear definition of this key parameter is fundamentally lacking.

Response. We apologize for the confusion. The original phrasing "detection probabilities span from 6.0 to 85.2 per 100 observations" was indeed ambiguous: 6.0 was a per-plant percentage (6.0%), not a count of 602 plumes. We have rewritten the affected lines (Section 4.2) using the standard percentage notation: "When adjusted for the number of satellite overpasses, the per-plant empirical detectability spans from 6.0% to 85.2% (median 20.4%; P10=10.0%, P90=63.7%)." To resolve the broader concern, the new "Definitions" paragraph at the start of Section 3 now distinguishes "detection" (the binary per-observation algorithm output) from "detectability" (the probability of detection given conditions, estimated either as a per-plant empirical fraction or via the machine learning model). These two are no longer interchangeable: detection is a 0/1 label, while detectability is a probability in [0,1].

38. Further minor 22

Reviewer comment. L287: remove the word 'from'.

Response. Fixed. Typo removed.

39. Further minor 23

Reviewer comment. Section 3.2: I believe that this text should come in a first dedicated subsection of the results (Section 4). Such a structure would make it easier for the reader to navigate Section 4. First present the performance of the model - in terms of the metrics introduced here - then go on to elaborate on these in connection with the key drivers of plume detectability.

Response. The closing paragraph of Section 3.4 (Plume Detectability Prediction) previewed results (which Table to focus on, why F1 is highlighted, and forward references to the appendix) that more naturally belong in the Results section. We have: Streamlined the Section 3.4 closing paragraph to retain only methodological content: the experimental design (item split vs. power plant split), the six classification metrics computed, and pointers to where results are reported. Moved the previously-previewed text (explanation of why F1, Precision, and Recall are highlighted, the structure of Table 3, and forward reference to the Appendix) to a new opening paragraph of the Model Performance subsection (Section 4.3) where it sits alongside the actual results.

40. Further minor 24

Reviewer comment. Section 3.2.2: In the last paragraph, a figure is being discussed with no reference to it leading the reader to wonder what is actually being talked about. Is this Figure 7?

Response. Fixed. The reference has been added: 'To visualize these local explanations on a global scale, we aggregate the plant-level SHAP results onto a grid (Figure D2).'

41. Further minor 25

Reviewer comment. L362: Round up/down numbers - detection is an integer value.

Response. Fixed. All detection counts have been rounded to integers.

42. Further minor 26

Reviewer comment. Section 4.3: There is an inadequate distinction made between the model performance (metrics depicted in Table 3) and plume detectability, which can be very confusing. This in part comes from the absence of any definition of plume detectability, but also the structure of the results section and the titles of the subsections. A subsection dedicated to the performance of the model would help, without tying this simultaneously to the key drivers of plume detectability. This would help disentangle results from discussion.

Response. The new "Definitions" paragraph in Section 3 (response to the previous comment) now formally separates "detectability" (the predicted quantity) from the model-evaluation metrics in Table 3 (accuracy, precision, recall, F1, AUC, kappa). In addition, we have renamed Section 4.3 to make its scope unambiguous: Old: "Meteorology, environment, and emissions enable accurate prediction of NO_x plume detectability" New: "Plume detectability is predictable with high accuracy across U.S. and global datasets" Section 4.3 now consists solely of model-performance content (Table 3 metrics, comparison across the All / Top-100 / Top-50 / Top-20 subsets, U.S. vs. global comparison). The discussion of which features drive detectability is contained separately in Section 4.4 ("Which features are most predictive of plume detectability?"), which deals with permutation importance and probability-of-detection curves. We hope this disentangles model performance from driver analysis as the reviewer suggested.

43. Further minor 27

Reviewer comment. Section 4.3: Given that the AUC score is a key metric in the evaluation and interpretation, it would be appropriate to see at least one ROC curve per analysis (one for US and one for Global), since this is the definition of AUC - area under ROC curve - with some text describing its result.

Response. We have added a new appendix subsection (Appendix D.1) with a figure showing the ROC curves on the held-out test split for the U.S. (hourly NO_x, "All") and global (annual NO_x, "All") models, together with the random-classifier diagonal. The accompanying text reports the test-set AUC values (U.S. 0.83, global 0.80) and confirms

they are consistent with the 5-run means (0.819 ± 0.006, 0.801 ± 0.003) in Table 4; we disclose that the plotted curves are the single best model (highest validation AUC) of five training runs per region.

44. Further minor 28

Reviewer comment. L389: Is this conclusion derived from comparing the different dataset scopes (All vs. Top-X emitters)?

Response. Yes. The conclusion compares separate models trained and evaluated on the four dataset scopes (All'', Top-100'', Top-50'', and Top-20''), as described in Section 3.3. We have revised the relevant paragraph in Section 4.1 to make this explicit: "Model performance improves when focusing on high-magnitude emitters, as seen by comparing the All'', Top-100'', Top-50'', and Top-20'' rows of Table 2, where each row corresponds to a separate model trained and evaluated on the indicated subset (Section 3.3). As the subset is restricted to higher emitters, Precision, Recall, and the F1 score all rise, with Precision showing the largest improvement. This trend suggests that on these subsets the model makes fewer false positive errors, since stronger emitters tend to produce less ambiguous plume signals."

45. Further minor 29

Reviewer comment. L405: These regions are notoriously difficult for satellite retrievals due to limiting factors such as high aerosol load, water vapour and cloud cover. This is most likely the reason for sparser training data in these regions.

Response. We agree that the sparser training data in parts of Asia, South America, and Africa is partly a consequence of conditions that make satellite retrievals more difficult, and we have revised the text in Section 4.2 to acknowledge this: "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."

46. Further minor 30

Reviewer comment. Section 4.4: I would bring this section forward to the first one of the results because it has bearing on all the other results sections (sec 4.2 & 4.3 at least). To illustrate this, take the result in Section 4.2 that there is a clear geographical distribution in detectability (L370), in which dry arid areas have more often plumes detected. This has likely to do with the fact that the surface albedo is high, leading to higher signal-to-noise is the satellite measurements. Looking at Figure 9, the surface albedo is indeed the second most important feature. Therefore, this geographical distribution can be explained by the feature importance, so a more logical flow of presentation would be to have the Section 4.4 before the others. Furthermore, the features that affect plume detectability is the most fundamental result.

Response. After careful consideration, we have retained the original section order because the geographic detection patterns in Section 4.2 are direct empirical outputs of the Automated Plume Detection algorithm (Section 3.2), upstream of the MLP that produces the performance and feature-importance results in Section 4.3 and Section 4.4. Reversing the order would place MLP-derived analysis before its empirical input, which we believe would obscure rather than clarify the methodological flow. To address the reviewer's underlying point (that the geographic patterns deserve explanation in terms of feature importance) we have made two additions: a forward reference at the end of Section 4.2 pointing readers to Section 4.4, and an interpretive sentence within the surface-albedo discussion in Section 4.4 that explicitly links back to the arid–humid contrast in Section 4.2. The explanatory narrative the reviewer asked for is therefore preserved, while the methodological dependency (algorithm output → MLP analysis) remains visible in the section order.

47. Further minor 31

Reviewer comment. L443: I think this surprising result needs more interpretation. It is somewhat questionable that plume detectability is higher for wind speed less than 2 ms-1 in light of what the literature says on this, but the authors give no satisfying explanation of why this would be. Wind speed should also show a hump shape much like solar zenith angle. Can it be that the real relationship is shielded by correlation with other features?

Response. The monotonic decrease in our results and the hump-shaped literature curves describe two different tasks: emission quantification requires a coherent plume structure that the retrieval can fit, which breaks down at very low wind, whereas detection asks only whether any NO_2 enhancement is present at the source, which is favored by low wind because NO_2 accumulates near the source rather than dispersing. We have added a sentence to the wind-speed paragraph in Section 4.4 making this distinction explicit.

48. Further minor 32

Reviewer comment. L444: Reference needed when referring to literature.

Response. We have added the citation (Bruno et al., 2024) directly after the 2–4 m/s claim: "We note that this is different from the ideal wind speeds for quantifying emissions, which prior literature shows to occur at moderate wind speeds (2–4 m/s) (Bruno et al., 2024)."

49. Further minor 33

Reviewer comment. L470: I think a more thorough presentation of the literature in Section 1 would help to substantiate this claim. This is given in the second paragraph of the introduction, however this could be expanded in more detail.

Response. Following the reviewer's suggestion, we have expanded the second paragraph of Section 1. We now layer the prior literature in three groups: (i) satellite NO_x quantification, OMI-era wind-rotated and EMG methods (Beirle et al., 2011; Liu et al., 2016; de Foy et al., 2015), TROPOMI extensions (Goldberg et al., 2019; Lange et al., 2022; Tang et al., 2024), and divergence-method catalogs (Beirle et al., 2019; Beirle et al., 2021; Beirle et al., 2023); (ii) the complementary plume-identification line, supervised NO_2 detection on

TROPOMI (Finch et al., 2022) and the more developed methane/CO_2 super-emitter literature (Kuhlmann et al., 2019; Cusworth et al., 2021; Cusworth et al., 2023; Lauvaux et al., 2022; Schuit et al., 2023; Rouet-Leduc and Hulbert, 2024; Dumont Le Brazidec et al., 2025; Liu et al., 2020; Bruno et al., 2024); and (iii) the theoretical and retrieval-side scaling that underpins detectability (addressed in our reply to L507 below).

50. Further minor 34

Reviewer comment. L515 Why was aerosol optical depth (AOT) not available? Aerosol is the most complex variable to accurately model in trace gas retrieval and therefore it is a very important factor in satellite data. I would expect to see AOT near the top of the feature importance in Figure 9, thus the absence of this parameter should be properly addressed in the text. This could also be one of the reasons limiting the AUC values (L387).

Response. We have revised Section 2.5 to introduce our aerosol treatment explicitly. We use the TROPOMI 354–388 nm UV aerosol index, which captures absorbing aerosols (e.g., dust, smoke, volcanic ash), an important aerosol-related interference in NO_2 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 NO_2 product, and incorporating it would require collocation with a separate product such as the CAMS reanalysis (Inness et al., 2019). We have therefore not included AOD in the present analysis and retain it as a direction for future work, which is now consistently reflected in both Section 2.5 and the limitations paragraph of Section 5. The revised Section 2.5 text reads: "The TROPOMI 354–388 nm UV aerosol index captures absorbing aerosols (e.g., dust, smoke, volcanic ash), which are an important aerosol-related interference in NO_2 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 NO_2 product and is not included in the present analysis."

51. Further minor 35

Reviewer comment. L507: If this is a main conclusion, and it appears so since half of the discussion centers around it, the theoretical expectations deserve an introduction in Section 1, with appropriate references to literature, to help substantiate this claim.

Response. We have added the theoretical and retrieval-physics background to Section 1 with appropriate citations. Specifically, the downwind column enhancement scales with emission rate and inversely with wind speed and plume spread (Varon et al., 2018); the effective NO_x lifetime is modulated by photolysis, OH concentration, and temperature (Valin et al., 2013; Romer et al., 2018; Laughner and Cohen, 2019); and retrieval-side sensitivity depends on air-mass-factor geometry, surface and cloud conditions, and background noise (Palmer et al., 2001; Boersma et al., 2011; Eskes and Boersma, 2003; Lorente et al., 2017; Verhoelst et al., 2021). These three threads are now explicitly motivated in the introduction, so the discussion in Section 5 builds on a framework the reader has already encountered.

52. Further minor 36

Reviewer comment. Appendix F: Title of this section is misleading. A detection to me implies the presence of a plume, whereas Section F1.2 deals with true negatives which means that there is no plume present. This leads to the question: how do the authors define a detection? From the rest of the subsection titles, this appears to mean correct classification, however this is nowhere explicitly stated.

Response. To remove the ambiguity, we have renamed Appendix H from "Case Studies in Automated Plume Detection" to "Case Studies in Algorithm Performance", and renamed its two subsections from "Successful Detections" / "Failed Detections and Their Causes" to "Correct Classifications" / "Misclassifications and Their Causes". The opening paragraph of Appendix H defines true positives, true negatives, false positives, and false negatives explicitly, which together with the new titles should make the categorization clear.

We thank both reviewers for the time they invested in this manuscript, and we hope the revised version meets their expectations.

Sincerely,

Ruizhe Huang and Sherrie Wang