Response to Reviewer #2

Manuscript: "Spatial resolution versus precision in CO_2 point-source emission retrievals from the DQ-1 spaceborne lidar: a plume-width-normalized interpretation"

Journal: Atmospheric Measurement Techniques (AMT)

Reviewer: Julia Marshall

We thank the reviewer for the careful and constructive assessment. The original sampling geometry, averaging definition, flux equations, Allan analysis, and use of inventory data all required substantive correction. We therefore reprocessed the controlled experiments and the unaveraged power-plant tracks with the measured 337.5 m sample-centre spacing. The revision uses signal-first averaging, fixed-anchor non-overlapping plume blocks, and a separate overlapping Allan calculation for local variability. It also separates the detection scale N_{det} from the emission-qualification scale N_Q; inventories are consulted only after both choices have been made.

For clarity, each point below follows the AMT response sequence: (1) reviewer comment, (2) author response, and (3) changes in the revised manuscript. Reviewer comments are reproduced verbatim in black italics; our responses are shown in blue; and manuscript changes are shown in dark red. Because line numbering may change during production, revision locations are given by section, equation, figure, and table number.

Point-by-point response

General comment: Overall assessment and relation to prior DQ-1 studies

Reviewer comment

This paper assesses the feasibility of estimating CO2 emissions from point sources, using vertically integrated, remote-sensing measurements of X_{CO_2}. The authors rightly point out that this capability is well-established for passive spaceborne sensors, but less so for lidar measurements. There are exceptions, however: both Zhang et al. (2025) and Han et al. (2024) have previously retrieved power plant emissions using spaceborne lidar data from DQ-1, and airborne lidar measurements have been used in various studies (e.g. Amediek et al., 2017; Wolff et al., 2021; Shi et al., 2023). The aspect that this paper aims to address is the appropriate spatial averaging length to improve the precision of the measurement without losing the ability to resolve the fine spatial structure. Unfortunately no conclusions can really be drawn – they show that changing the averaging length changes the retrieved emissions, and then apparently adjust this length in a somewhat arbitrary manor to best match inventory values. This is not how data-driven emission estimation is supposed to work.

The approach taken in the current study is quite different to that of Han et al. (2024), though they are using the same data. The approach is much closer to that of Zhang et al. (2025), though this paper (or rather the preprint, Zhang et al. (2024b)) is only cited in passing as a "validation study" to confirm the accuracy of the measurement. However, this is not at all the focus of Zhang et al. (2025), perhaps the authors meant instead to cite Zhang et al. (2024a)?

In any case: there are two published studies that have already looked at the retrieval of power plant emissions from DQ-1 data (Han et al., 2024; Zhang et al., 2025), but there is essentially no discussion of how this paper's approach differs. It is a pity that the authors did not engage a bit more with these studies, as they might have avoided some serious methodological errors, which essentially invalidate the results. These issues are outlined in more detail below. Given the

impact of these errors on the results and conclusions, the paper cannot be considered for publication without major revisions.

Author response

We sincerely appreciate the valuable suggestions you have provided and deeply regret that the initial draft did not meet your expectations. In preparing the revised manuscript, we have carefully considered and incorporated each of your valuable suggestions. In the original manuscript, the averaging scale was not clearly separated from the external inventory, so the procedure could be read as inventory tuning. In the revised analysis, no EPA CEMS or Climate TRACE value enters candidate screening, plume fitting, quality scoring, N_{det}, or N_Q. N_Q is a satellite-only qualification optimum, not a truth-based accuracy optimum. The Introduction now discusses Han et al. (2024) and Zhang et al. (2025) directly: those studies established the feasibility of DQ-1 point-source retrievals, whereas the present study asks how the averaging support should be selected.

The controlled experiments give a width-normalized detection relation, Ldet/W = 1.209 (95 % confidence interval: 1.158–1.259), over 60 stability–distance geometries. The observational detection branch contains 48 complete six-scale cases. The separate flux branch contains 44 pre-screened cases, of which 38 retain at least one eligible scale. The full 38-case set is now supplied as Supplement Table S1; Table 3 remains a compact 20-case illustration. In the complete set, NQ = 5, 10, 15, 20, 25, and 30 occurs in 25, six, three, two, one, and one cases. Among 35 cases with both scales, NQ is coarser than Ndet in 10, equal in 13, and finer in 12. We have therefore removed population inference from the selected 20-row display subset.

The practical conclusion is that one averaging factor should not be imposed on both tasks and that the two optima are not universally ordered. The observed maximum of D(N) determines Ndet, with the controlled width relation used only to initialize the search. NQ is then selected independently from the same prespecified satellite-only gates at every candidate scale. EPA hourly CEMS and Climate TRACE monthly values are joined after selection as external context. We no longer describe the EPA values as transport-time matched because the single cross-plume transect and track-normal Ueff do not uniquely determine an along-plume travel time.

For every real overpass, one fine-resolution plume centre is fixed first. The calibrated online and offline signals are recomputed in outward, non-overlapping blocks for N = 5, 10, 15, 20, 25, and 30, followed by retrieval and cross-sectional-flux calculation. Every eligible scale must retain at least 15 profile blocks, including at least five on each side of the centre, and pass the same fit, geometry, uncertainty, containment, and background gates. After NQ is fixed, adjacent-scale agreement is reported separately; 11 of 38 cases meet the factor-of-two stability test.

The validation structure follows the aspects of the CHARM-F/MERLIN analysis of Wolff et al. (2021) that are identifiable for single DQ-1 transects: background and containment checks, known-input simulations, adjacent-scale flux consistency, and plant- and track-level clustering. CHARM-F-specific distance and turbulence thresholds are not transferred to DQ-1, and the inventory comparison remains the final, post-selection step.

The correction from 70 m footprint size to 337.5 m sample-centre spacing required a full recalculation, not a relabelling of axes. The former 50–100- and 75–150-point recommendations have been withdrawn. The revised field analysis uses N = 5–30 (1.6875–10.125 km), reconstructs disjoint signal-first blocks from the native records, and fixes both scale decisions before inventory values are attached. The controlled result is now reported as a detection-scale relation; the real observations provide support for the two-stage workflow without establishing a universal field coefficient or a truth-based emission optimum.

Changes in manuscript

Location in revised manuscript: Abstract; Introduction; Methods 2.4–2.5; Results 3.2–3.3; Figs. 13–18; Tables 1–3; Discussion 4.1–4.3; Conclusions

The central question is now stated as two related but non-identical objectives. N_{det} maximizes observed plume enhancement relative to local scale-matched variability and is used to detect and anchor the plume. N_Q is selected independently from fixed-anchor block inversions that satisfy prespecified fit, geometry, uncertainty, containment, and background-cleanliness conditions; adjacent-scale agreement is then reported separately. The approximate detection-width classes are defined in Sect. 2.5.2. Table 1 instead summarizes the selection rules, required satellite diagnostics, and reporting outputs for N_{det} and N_Q. The observed N_Q frequencies are reported as results, not used as prespecified table thresholds, and neither scale is selected by agreement with an inventory.

The controlled experiments support $L_{det}/W \approx 1.2$ as a first-order model scaling. The observations support a two-stage scale-selection workflow rather than a fixed offset between N_{det} and N_Q. In the complete 38-case set, 31 cases select N_Q = 5–10 and seven select N_Q = 15–30. The ordering relative to N_{det} is mixed, so the revised conclusion is an independent, evidence-constrained decision at each stage rather than a claim that emission qualification always needs more averaging.

Major comment 1: Spatial sampling distance and averaging geometry

1.1 Footprint size and sample-centre spacing

Reviewer comment

Spatial sampling distance

The most serious concern relates to the assumptions about the spatial sampling distance of the instrument. The authors use 70 m throughout, which is the dimension of the footprint of a single shot measurement. Each shot pair is separated by 350 m, as can be seen in e.g. Fig. 1 of Han et al. (2024). (Zhang et al. (2025) give the footprint spacing as only 330 m.1) This is a fundamental mistake, and affects most of the results in the study. At very least, it changes all of the results in section 3.1 (Figs. 3–12).

Author response

We sincerely apologize that the difference between 70 m and 337.5 m was not taken into account when preparing the initial draft, which led to a problem that could have been avoided. In the revised manuscript, 337.5 m is used as the native sample-centre spacing in every conversion from N to physical support. The approximately 70 m value is retained only as the along-track footprint dimension. All simulation distances, plume-width scaling, observational block lengths, and figure annotations have been recalculated accordingly.

This correction changes the physical interpretation of the averaging substantially. The six observational scales N = 5, 10, 15, 20, 25, and 30 correspond to nominal supports of 1.6875, 3.375, 5.0625, 6.75, 8.4375, and 10.125 km, respectively. The revised manuscript therefore discusses kilometre-scale support and the number of independent block values retained across a plume. The 20 Hz pulse-repetition frequency is reported as an instrument property, but the analysis uses the observed sample-centre geometry rather than inferring the along-track spacing from the footprint or PRF.

The requested numerical comparison is particularly clear for N = 15. In the manuscript originally reviewed, N = 15 was interpreted as 15 × 0.070 km = 1.05 km. In the revised analysis it is 15 ×

0.3375 km = 5.0625 km (reported as 5.06 km). The physical support is therefore 4.82 times larger than originally stated. The same factor applies to every conversion: N = 5 changes from 0.35 to 1.6875 km, N = 30 from 2.10 to 10.125 km, and the former N = 150 example from 10.5 to 50.625 km.

Effect on the original analysis. This correction cannot be handled by changing tick labels because the number of independent samples spanning a plume also changes. At a fixed plume width, a 5.06 km block suppresses the peak more strongly and produces fewer cross-plume observations than a 1.05 km block. We therefore regenerated the sampled plume profiles, repeated signal-first block averaging, repeated the forward fits and Monte Carlo summaries, recalculated the width-normalized detection relation, and rebuilt the real-overpass cross-sectional-flux profiles. Figures 3–18 and Tables 1–3 refer only to this recalculated analysis.

Effect on the conclusions. The old numerical ranges of 50–100 and 75–150 points are no longer conclusions of the paper. The revised upper observational test is N = 30 (10.125 km). The same support rule is now applied at every scale: at least 15 profile blocks in total and at least five on each side of the fixed centre. N = 30 is selected in one of the complete 38 cases, but it remains diagnostic for Luohuang because that particular profile contains only 13 blocks.

Averaging factor N	Nominal support N × 0.3375 km	Interpretive role in the revised study
5	1.688 km	Fine end of the observational detection search
10	3.375 km	Median observed N_{det} among complete cases
15	5.062 km	Intermediate simulation and profile-search scale
20	6.75 km	Coarser profile support
25	8.438 km	Median observed maximum-R^2 profile-agreement scale
30	10.12 km	Upper tested observational scale

Changes in manuscript

Location in revised manuscript: Methods 2.1.1, 2.3–2.5; Results 3.1–3.3; Figs. 3–18; Conclusions

All physical supports used the 337.5 m sample-centre spacing; the approximately 70 m laser footprint was not used to convert N to distance.

Major comment 2: Block averaging and the apparent continuous tracks

Reviewer comment

I cannot understand how the authors could have analyzed actual DQ-1 data (in Sects. 2.3 and 3.2, for example) without having realized the inconsistency between their 70 m assumption and the true sampling at some point. If the DQ-1 soundings were really averaged as explained in Fig. 13, there should no longer be a continuous-looking line of measurements for higher numbers of points included in the averaging. Compare to Figs. 4b or 8: The spacing of the plume cross-section changes as a function of the number of points that are averaged, as one would expect. If we consider the case of averaging 150 points for the overpasses shown in Fig. 13, given the actual spacing of 350 m between each measurement, there should only be one point per 52.5 km.

This averaging would result in only two to three points for the Majuba plots, and perhaps three to four for the Lalitpur plots, but in Fig. 13 we see a continuous measurement track.

How can this be? The only thing I can think of is that a moving averaging window was used, essentially reusing all the measurements up to 150 times. This is statistically problematic, but also does not explain why the path length for the Majuba plots is the same for all averaging lengths: if 150 shot-pairs have to be averaged together, even allowing for the inappropriate use of a moving average, the first 74 shots (≈ 26 km) should be truncated. This would be a visible difference in the scale shown in Fig. 13.

Notes

1 Reviewing this paper led me to read the results inZhang et al. (2025) as well, to see how they dealt with the averaging/spatial resolution question. Their results also seem to be fundamentally inconsistent with the spatial resolution of DQ-1. Consider Fig. 4 of that paper: the text states that there are 76 points that were designated as "in plume", shown in blue in Fig. 4d, along with the modelled emission plume (green and red dots). This modelled emission plume is apparently sown in Fig. 4c, and the position where the measurements cross the plume appears to be ≈ 10 km wide. However, 76 points should cover a distance of around 26 km. Is the model scale completely wrong? Are these even real measurements? The fact that the DQ-1 data have still not been publicly released makes it impossible to check.

Author response

We agree that the original figure and description made the averaging procedure ambiguous and could give the impression that a moving window preserved an artificially continuous plume track. We therefore returned to the original unaveraged Level-2C sequence and rebuilt every quantitative plume profile. For each overpass, the fine-scale downwind maximum is identified once and retained as a common anchor. At every tested N = 5, 10, 15, 20, 25, and 30, complete non-overlapping N-sample blocks are then constructed from that anchor toward both ends of the valid track. Calibrated online and offline signals are summed within each block before the nonlinear XCO_2 retrieval is evaluated. Consequently, the values entering the plume fit are based on disjoint native samples, the number of profile values decreases as N increases, and incomplete edge blocks are not used.

One-sample-shifted AVS windows are used only for the local plume-excluded Allan calculation; they are never passed to the flux inversion as independent profile values. Figure 15 makes the distinction visible. Its first column shows the plant, native valid track, selected non-overlapping AVS blocks, ERA5 vectors, and the track-normal U_{eff}. The other columns show the six-scale flux and uncertainty, fit and background diagnostics, and the profile fit at the selected N_Q. The reported U_{eff} values are 0.58, 1.50, 2.74, 2.14, and 7.81 m s−1.

Comparison with the original Fig. 13. The eight panels for each of Majuba and Lalitpur showed apparently continuous coloured tracks even at N = 100–150. As the reviewer noted, a true N = 150 non-overlapping average at the corrected spacing has a nominal support of 50.625 km and cannot produce a dense sequence over those map extents. That presentation has been removed. It is not used as evidence in the revised manuscript, and the associated case-specific conclusion that 75–150 points was preferred is also removed.

Luohuang illustrates the consequence in the real data. The selected N = 25 result contains 15 retained blocks, whereas the N = 30 diagnostic contains only 13 and is excluded by the minimum-support requirement. This exclusion is made before the Climate TRACE value is joined. The example is included because it lets the reviewer see how a nominally smoother profile can become less defensible once independent support is counted explicitly.

Figure 15 replaces the former map sequence with a four-part audit for each of five overpasses: map and wind geometry, the six-scale flux results, scale-specific fit and background diagnostics, and the fitted profile at the selected N_Q. It therefore shows both why a scale was retained and why neighbouring scales failed.

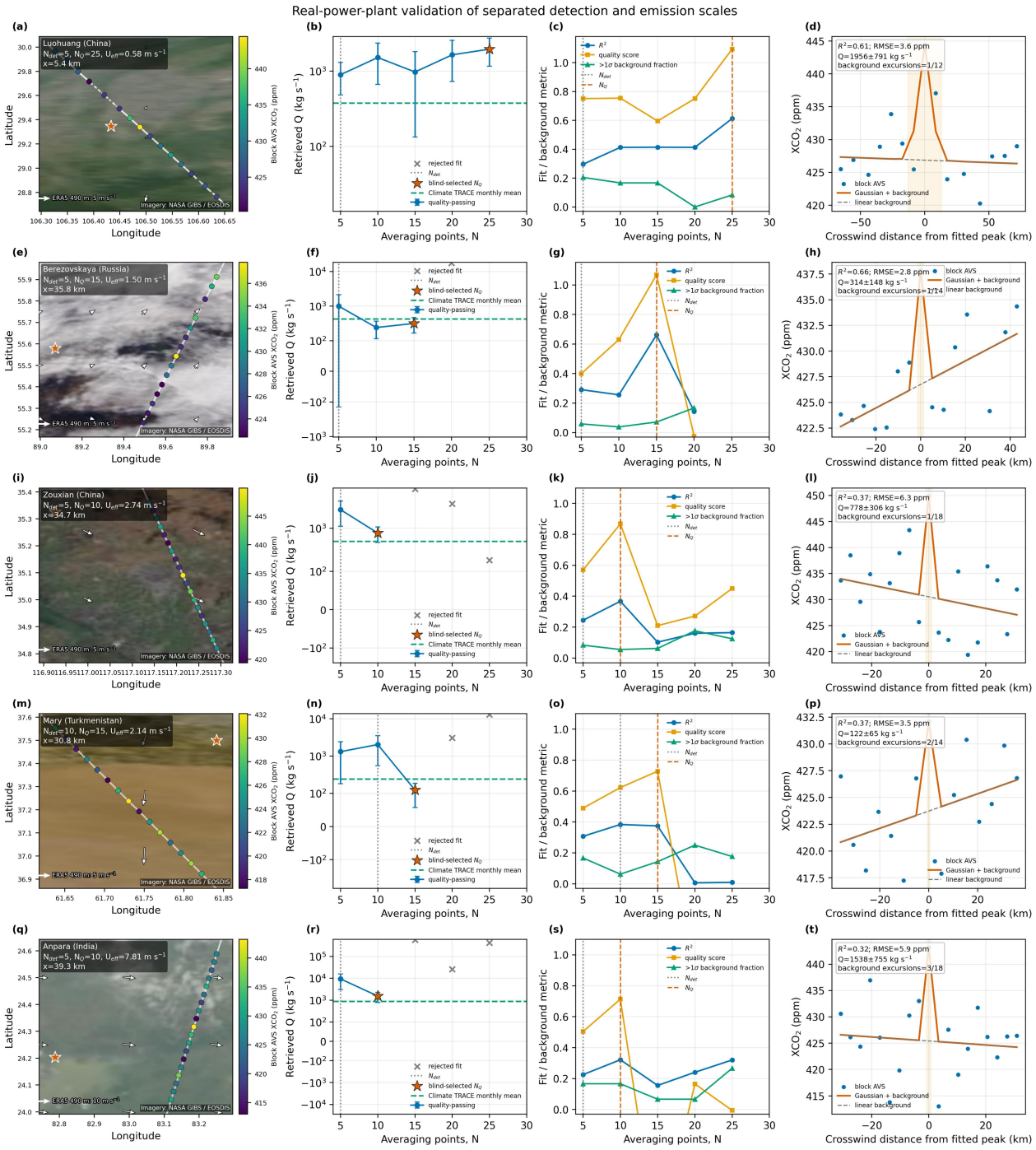

Figure R1. Revised manuscript Fig. 15. Reference-emission-blind six-scale block-averaging analysis for Luohuang, Berezovskaya, Zouxian, Mary, and Anpara. The tightly framed maps retain the plant, native valid track, selected non-overlapping signal-first AVS blocks, and sparse ERA5 wind vectors. Each map reports the track-normal U_{eff} used in the flux calculation. The remaining columns show $\hat{Q}(N)$, the prespecified fit/background diagnostics, and the selected profile fit. External values are attached only after N_Q is frozen.

Changes in manuscript

Location in revised manuscript: Methods 2.5.1–2.5.2; Results 3.3.1; revised Fig. 15 and its caption

The revised manuscript no longer displays a moving-average plume track. Quantitative plume fitting and emission retrieval use only fixed-anchor, non-overlapping signal-first AVS blocks; the separate one-sample-shifted windows are confined to the local-Allan estimator.

Major comment 3: How spatial averaging changes plume information

Reviewer comment

The whole focus of this study is finding the optimal spatial averaging length to best retrieve small-scale features like plumes. This is a real limitation of current IPDA lidar technology. While the repetition rate is sufficient to resolve emission plumes from aircraft, faster-moving spaceborne platforms present additional challenges. This is made clear in several of the studies cited by the authors, such as this passage from Kiemle et al. (2017):

Plume detection from quickly moving platforms such as satellites, however, requires high spatial resolution, which will only be achievable with a small laser footprint, a high laser pulse repetition frequency (PRF) and a single- measurement (non-averaged) retrieval.

Note that Kiemle et al. (2017) assume a PRF of 500 Hz for this feasability study, compared to the 20 Hz of DQ-1.

Author response

The revised controlled analysis evaluates averaging before any inventory comparison. Non-overlapping block averaging reduces scale-dependent variability but simultaneously lowers the plume peak and leaves fewer independent samples across the plume. The resulting response is finite rather than monotonic: intermediate support can improve profile stability, whereas coarse support eventually flattens the plume and produces degenerate fits. This is also the distinction emphasized by Kuhlmann et al. (2019), who showed that spatial averaging can expose weak plume signals but risks diluting their edges and that plume detectability is not itself an emission-accuracy criterion.

The comparison with Kiemle et al. (2017) is now explicit in Discussion 4.1. Their 500 Hz feasibility concept is intended to retain dense single-measurement sampling from a fast-moving platform. DQ-1 operates at 20 Hz, and the Level-2C samples analysed here have a 337.5 m along-track sample-centre spacing. Thus N = 5 already corresponds to 1.69 km of nominal support and N = 20 to 6.75 km. The revised manuscript therefore treats averaging as a finite precision–resolution balance rather than an unrestricted path to improved precision.

We quantified the balance support relative to the Gaussian-plume FWHM W. The baseline audit gives $L_{det} = 1.182W$ ($R^2 = 0.968$), and the extension to 60 unique stability–distance geometries gives $L_{det} = 1.209W$ with a two-way-bootstrap 95 % interval of 1.158–1.259 ($R^2 = 0.971$). The pooled coefficient is used consistently for initialization and field transfer. Recalculation gives 19/46 exact agreements, 35/46 agreements within five samples, and 22/46 predictions in the 5 %-near-optimal set; the association remains non-significant (Spearman $\rho = 0.071$, $p = 0.641$). These bins initialize the search only; final observational N_{det} remains the maximum of D(N).

For emission reporting, N_Q is selected as the eligible maximum of the reference-emission-blind score, with the stated 0.05 tie-break. Supplement Table S2 now gives a direct decision-rule sensitivity analysis. The choice is unchanged in all 38 cases when the inverse-uncertainty penalty is removed, the uncertainty weights are halved or doubled, or the near-tie tolerance is widened to 0.10; removing the precision reward or setting the near-tie tolerance to zero changes one case each. Strict gates retain

36 cases and preserve 35 choices. Adjacent-scale agreement is evaluated separately, and 11 of 38 cases meet the factor-of-two criterion.

Compared with the manuscript originally reviewed, the revision no longer stops at the qualitative sequence of noisy fine scales and over-smoothed coarse scales, nor does it use proximity to an external emission value to identify the preferred N. It now reports a controlled L_{det}/W relation, a reference-emission-blind N_Q decision, and a separate adjacent-scale stability test.

The five detailed overpasses demonstrate the coarser-qualification regime without defining it by inventory agreement. Luohuang changes from $N_{det} = 5$ (1.69 km) to $N_Q = 25$ (8.44 km); Berezovskaya from 5 to 15 (1.69 to 5.06 km); Zouxian from 5 to 10 (1.69 to 3.38 km); Mary from 10 to 15 (3.38 to 5.06 km); and Anpara from 5 to 10 (1.69 to 3.38 km). Each selected scale is obtained before an external emission is attached, and the full N sequence, including failed scales, is shown in Fig. 15.

These five examples are deliberately presented as mechanisms for the coarser-support regime, not as its frequency in the archive. The population result now comes from all 38 eligible cases in Supplement Table S1. Among the 35 paired cases, NQ is coarser than Ndet in 10, equal in 13, and finer in 12. Thus, detection and flux qualification are different objectives, but neither a fixed offset nor a universal ordering is supported.

The practical statement added to the Discussion is consequently conditional but usable. Begin at N = 5 after fixing the plume centre. Use the width relation only to initialize or prioritize the detection search, and retain the observed maximum of D(N) as N_{det}. For flux qualification, move beyond N = 5 only when the same anchored, signal-first reprocessing gives a material improvement in the prespecified score and still satisfies support, geometry, containment, background, and formal-uncertainty gates. Report the flux as scale-stable only if an adjacent tested N agrees within a factor of two. This is the main methodological conclusion produced by the revision.

The background-cleanliness metric contributes information that R^2_prof alone cannot provide. It counts unexplained excursions outside the fitted plume region relative to the scale-matched local variability. A large count indicates that a convincing central Gaussian coexists with unrelated structure elsewhere on the profile. Together with containment and minimum-support checks, this prevents the score from rewarding smoothing alone. The metric is calculated without an inventory and is shown for the detailed real cases.

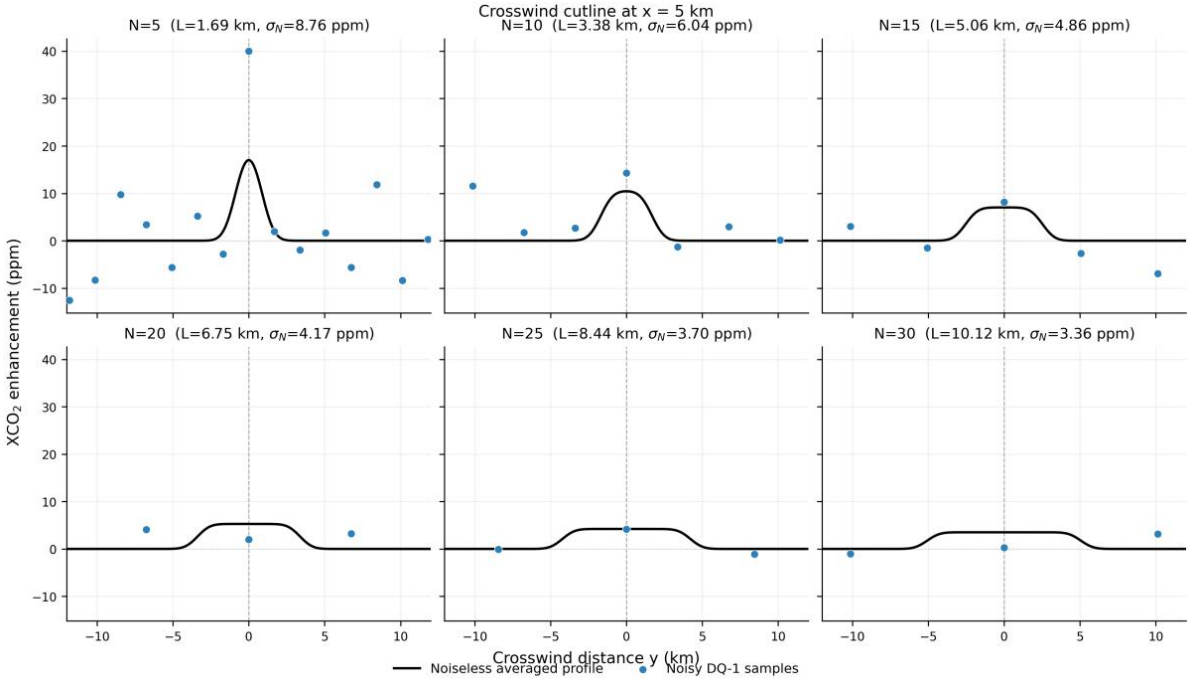

Figure R2. Revised manuscript Fig. 4. Crosswind profiles after non-overlapping block averaging at N = 5–30. The figure shows the simultaneous decline of random variability and plume contrast as support becomes coarser.

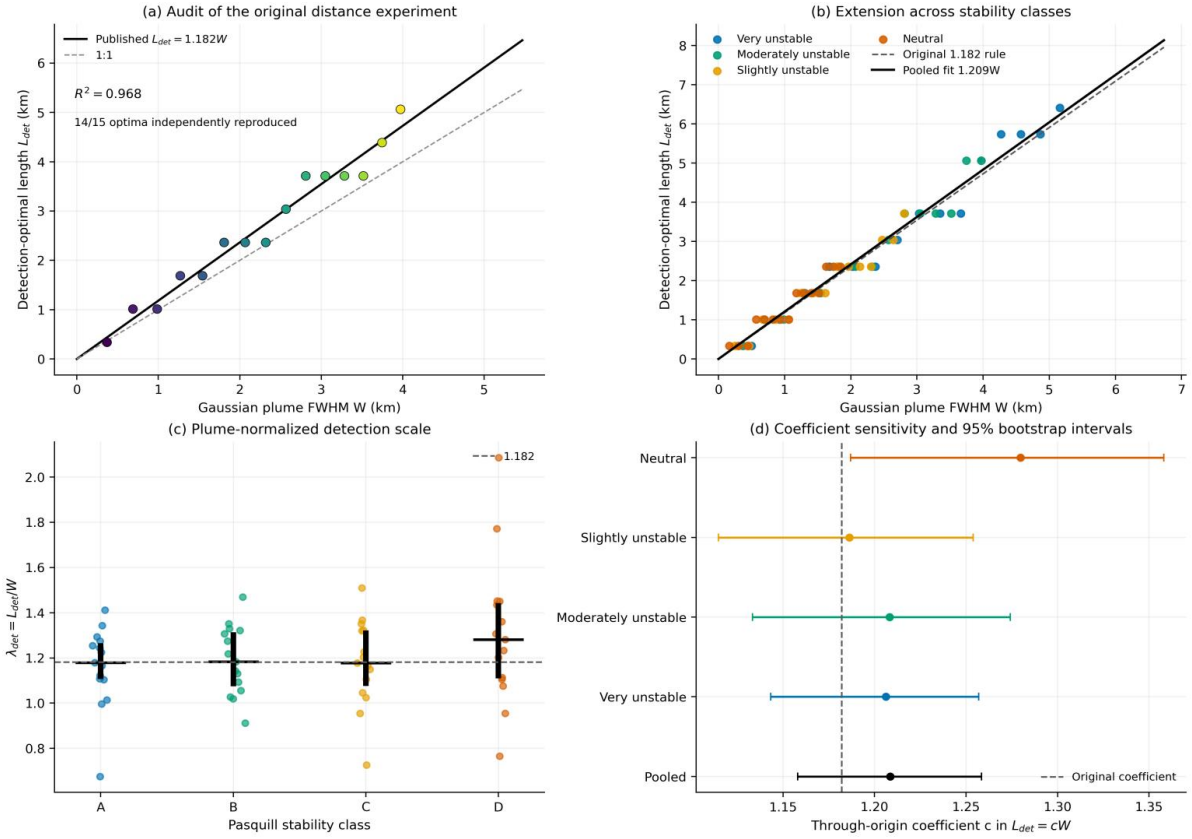

Figure R3. Revised manuscript Fig. 13. Plume-width normalization across stability classes and bootstrap uncertainty of the through-origin coefficient. The relation is reported as a controlled-simulation prior, not a universal observational law.

Changes in manuscript

Location in revised manuscript: Methods 2.5 and Table 1; Results 3.1–3.3; revised Figs. 3–15; Discussion 4.1–4.3; Conclusions

The revision now distinguishes a plume-width-normalized detection balance from an independently qualified emission scale. The field data support two detection classes, N_{det} = 5–10 and N_{det} = 20–30, but no independent distance or wind-speed threshold; therefore the observed D(N) maximum, not a fixed environmental cutoff, controls the final detection scale. Five real overpasses reproduce the coarser N_Q regime predicted by the simulations, while the full accepted set shows that its magnitude and ordering are scene dependent.

Major comment 4: Emission-estimation method

4.1 Equation attribution and model formulation

Reviewer comment

Emission estimation method

The description of the emission estimation method is insufficient, erroneous, and seems to have been copied verbatim from a previous study without citation.

Beginning with Eq. 1, the authors cite Bovensmann et al. (2010), but this formulation does not match the relevant equations in the source paper, some combination of their Eqs. A1, A2, and A4. To match Eqs. A1 and A2 in Bovensmann et al. (2010), should be replaced by simply x. In this reference paper, x_0 is only introduced in Eqs. A4 and A5 to simulate emission sources with finite cross-sections, rather than point sources. In the manuscript under review, x_0 is never introduced.

Likewise, at first I struggled to decipher how Eq. 3 was derived from Eqs. 1 and 2, until I realized that the a in Eq. 3 was not the same a as in Eq. 1, nor is it ever defined. The variable μ was never introduced, and it seemed like an error to have b in both the first and second terms of Eq. 3. I consulted what seems to be the most relevant formulation in the cited paper (Eq. 1 in Reuter et al. (2019)), and it really did not relate to the derivation in the current paper.

Finally, I found the formulae (complete with some of the same errors) in Zhang et al. (2025). In fact, the majority of the text in Sect. 2.2 of the current manuscript was identical to that of Zhang et al. (2025). Given the "typo" in line 118, where the number 130 appears for no reason, I decided to check the preprint version of Zhang et al. (2025), Zhang et al. (2024b), which the authors did cite on line 58, but in a different (and incorrect) context. There I found the text in question, taken directly from Sect. 2.2, around line 130 of that paper. The preprint cited Bovensmann et al. (2010) in the introductory sentence of their Sect. 2.2 as well, but this was updated prior to final publication. The extra (first) b in Eq. 3 should have been a subscript, and the careless copy-paste explains some other formatting errors related to superscripts (e.g. in line 115). The manuscript under review kept the unit error for the variable w: it is given as kg m2 on line 135 of Zhang et al. (2024b) and on line 121 of the current paper, but was corrected to kg m—2 before publication in Zhang et al. (2025).

This is also flagged in the automated similarity report, though it appears less grievous as part of it shows up as copied from Zhang et al. (2024a) while the other part shows up as identical to Zhang et al. (2025). It is also relevant to point out that there is (only) one overlapping co-author between the manuscript under review and Zhang et al. (2025) – the second author of this paper is the seventh author of Zhang et al. (2025).

Author response

We recognize that the original Sect. 2.2 may have had some methodological and attribution issues. Its wording and notation followed Zhang et al. (2024b, 2025) rather closely without adequately explaining the relationship to that work, and we sincerely apologize for the insufficient attribution. Section 2.2 has been substantially reconstructed from the primary sources, with the earlier DQ-1 studies cited for their actual roles and the present averaging-scale contribution presented separately.

Equation (1) is now identified as the vertically integrated Gaussian point-source formulation. The finite-source offset x0 is not used. The revised manuscript states the actual three-dimensional dispersion implementation, $\sigma k(x) = \exp[Ak,0 + Ak,1 \ln x + Ak,2(\ln x)2]$ for $k \in \{y, z\}$, and Supplement Table S3 lists both crosswind and vertical coefficients with their units. Vertical integration yields the column form in Eq. (1). This replaces the erroneous finite-source ratio and removes ambiguity about the implemented dispersion law.

The estimator is now unambiguous. Equation (4) is restricted to controlled simulations, where Q is fitted for prescribed wind, stability, distance, and sampling. Real DQ-1 overpasses use Eqs. (5a)–(5b): the fitted Gaussian enhancement is integrated and multiplied by U_{eff} and the dry-air column conversion. The two analyses share a Gaussian physical framework but use different estimators. Table 3 therefore reports fluxes obtained from the fixed-anchor, signal-first cross-sectional procedure, not from the ambiguous least-squares description in the original manuscript.

Compared with the original method, the undefined finite-source term, reused symbols, and incorrect water-column unit have been removed. Equation (3) is now an empirical Gaussian-plus-background profile fit, Eq. (4) is limited to controlled simulations, and Eqs. (5a)–(5b) define the real-overpass cross-sectional-flux estimate.

Changes in manuscript

Location in revised manuscript: Introduction, paragraphs distinguishing the present averaging-scale study from the earlier DQ-1 emission studies; Methods 2.2, Eqs. (1)–(8), variable definitions, flux fitting, fit metrics, and wind geometry. The revised manuscript no longer presents text or notation adapted from earlier work without identifying the relevant source.

The finite-source offset x_0 introduced by Bovensmann et al. (2010) for sources with a non-zero initial cross-section is not used here because the power plants are represented as point sources.

Equation (3) is used to characterize the observed plume and background; it is not derived from Eqs. (1) and (2).

4.2 Flux fit and interpretation of R^2

Reviewer comment

Besides these problems, the method section does not sufficiently explain how they solved for the flux rate, F. Which parameters had to be fit? Which values were compared to calculate the R2 values that are quoted throughout? It is not possible to get a negative R2 value, despite what is claimed in the text. (Could it be that the numbers given are rather the Pearson correlation coefficient?)

Author response

We have rewritten this part of the method because the original version used the same symbol R^2 for two different diagnostics and did not distinguish the simulation estimator from the real-overpass estimator. Equation (3) describes the observed cross-plume profile with a linear background and a Gaussian enhancement. For the controlled simulations only, trial values of $\hat{Q}$ are propagated through Eqs. (1)–(2) at the retained simulated locations, and Eq. (4) obtains $\hat{Q}$ by minimizing the forward-

model residual sum of squares. This is a one-parameter controlled forward fit for each prescribed wind, stability, distance, and averaging configuration. For real DQ-1 overpasses, Eq. (4) is not used. The fitted Gaussian enhancement is converted instead with the cross-sectional-flux estimator in Eqs. (5a)–(5b), $\hat{Q}_{CSF} = 10^{-3} U_{eff} \int \Delta mCO_2 obs(s) ds = 10^{-6} (MCO_2/Mair)[(\bar{p}s - \bar{w}g)/g] U_{eff} \hat{A}\hat{\sigma}\sqrt{(2\pi)}$, reported in kg s−1. Thus, the two estimators share the same Gaussian physical framework and dry-air column conversion, but their estimators are explicitly distinct.

The revision now keeps the two coefficients of determination separate. For the controlled simulations in Figs. 5–12, forward-model agreement is denoted R^2_flux and compares the noisy/block-averaged observed enhancement with the enhancement predicted by the fitted Gaussian-plume forward model:

$$R^2_{\text{flux}} = 1 - \frac{\sum_{i \in \Omega} \left(\Delta X_i^{\text{obs}} - \widehat{\Delta X}_i^{\text{mod}} \right)^2}{\sum_{i \in \Omega} \left(\Delta X_i^{\text{obs}} - \overline{\Delta X}^{\text{obs}} \right)^2}$$

This is the quantity for which negative values occur in the simulation figures. Because the constrained forward plume template need not outperform the sample-mean reference, its residual sum of squares can exceed the total sum of squares and R^2_flux can be negative. These values therefore indicate an uninformative forward fit; they are not Pearson correlation coefficients and are not treated as successful retrievals.

For the observation-based scale analysis, we use a different diagnostic, R^2_{prof}, calculated from the Gaussian-plus-linear-background fit in Eq. (3):

$$R^2_{\text{prof}} = 1 - \frac{\sum_i \left(y_i - \hat{y}_i^{\text{G}} \right)^2}{\sum_i \left(y_i - \overline{y} \right)^2}$$

Here $\hat{y}_i^{G}$ is the fitted Gaussian-plus-background profile value. N_{R^2} is retained only as a population-level profile-agreement diagnostic. It is not used as an emission optimum. For the detailed real cases, N_Q is selected from the complete prespecified satellite-only quality score after geometry, uncertainty, containment, and background-cleanliness gates are applied. It is an emission-qualification optimum within that decision rule, not a truth-based emission-accuracy optimum. This prevents ordinary R^2 smoothing behaviour from defining the reported emission scale by itself.

The revision now states exactly which values enter each calculation. In a controlled realization, the block-averaged simulated enhancements are the observations and the plume values generated for trial Q are the predictions. Wind, stability class, downwind distance, and sampling support are prescribed. Minimizing the residual sum of squares therefore estimates one source-strength parameter for that realization. Repetition over the prescribed noise ensemble produces the reported mean behaviour, bias or RMSE, and Monte Carlo spread.

For a real profile, the fitted parameters are the Gaussian enhancement amplitude $\hat{A}$, plume-centre coordinate, Gaussian width $\hat{\sigma}$, linear-background intercept, and background slope. These parameters produce R^2_prof and the Gaussian area $\hat{A}\hat{\sigma}\sqrt{(2\pi)}$. The latter, rather than a trial-Q forward fit, is passed to the cross-sectional-flux equation. This list of fitted and prescribed quantities was not explicit in the manuscript originally reviewed and is now given immediately after the relevant equations.

The original use of a single R2 label across simulations and observations has been withdrawn. R2flux supports the controlled fit diagnostics, whereas R2prof describes real-profile agreement. The maximum-R2 profile scale is retained only as a diagnostic population summary; it is not called an emission optimum. NQ is reported only after all independent gates and the reference-emission-blind score are applied.

Table 3 is designed to expose that decision. It gives plant and overpass identity, N_{det} where available, N_Q, retrieved flux and formal uncertainty, R^2_{prof}, RMSE, U_{eff}, source distance, the >1σ background-excursion metric, and the post-selection external comparison. Retained-block counts, wind magnitude, crossing geometry, and adjacent-scale stability are documented in the surrounding text, Fig. 15, and the source-data package rather than presented as separate columns in Table 3. These fields allow the reader to understand the selected scale without inferring it from the inventory ratio.

Changes in manuscript

Location in revised manuscript: Methods 2.2, Eq. (3), the controlled-simulation Eq. (4), the real-overpass cross-sectional-flux Eqs. (5a)–(5b), the diagnostics and uncertainty in Eqs. (6)–(8), Results 3.1 and 3.3, revised Figs. 5–12, and Fig. 16.

R^2_{flux} is defined for the controlled forward emission fit, whereas R^2_{prof} describes the real Gaussian-plus-linear-background profile. Neither metric alone defines N_Q. The complete 38-case set selects N_Q = 5–10 in 31 cases and N_Q = 15–30 in seven cases, including one eligible N_Q = 30 case. Coarser support is therefore allowed at every tested scale but must pass the same minimum-support and quality gates and improve the reference-emission-blind score. These are qualification decisions, not truth-based accuracy optima.

4.3 Uncertainty and wind geometry

Reviewer comment

There is also no explanation of how the uncertainties on the emission estimates are calculated. What are the leading error terms? Looking at the comparisons to real data in Fig. 13, the wind direction seems quite far from perpendicular to the flight track. How was this dealt with?

Author response

We now separate both the estimator and the uncertainty treatment used in the controlled simulations from those used for individual real profiles. In the simulations, Eq. (4) directly fits Q with the Gaussian-plume forward model, and the quoted uncertainty is the standard deviation of $\hat{Q}$ over the prescribed noise realizations; the wind direction is fixed at 270° and perpendicular to the idealized satellite track, so U_{eff} = U by construction. For real DQ-1 overpasses, the fitted Gaussian area is converted with the cross-sectional-flux estimator in Eqs. (5a)–(5b), $\hat{Q}CSF \propto \hat{A}\hat{\sigma}U_{eff}$. Equation (8) therefore propagates the fitted enhancement amplitude, fitted width, and effective-wind term uU_{eff}/U_{eff} in quadrature. This remains a formal fit uncertainty, not a complete error budget: the dry-air conversion factor is treated as fixed, and background-model error, amplitude–width covariance, plume rise, stability, unresolved wind-direction error, and other transport errors are not fully represented.

The flight-track angle is handled differently in the controlled simulations and in the real DQ-1 cases. In all controlled simulations, the wind direction is prescribed as 270° and is perpendicular to the idealized satellite track. The simulation wind speeds quoted throughout the paper are therefore the prescribed magnitudes U, with U_{eff} = U, and no additional wind projection is applied. For real overpasses, the track can cross the plume obliquely, so the emission calculation instead uses the effective wind speed normal to the local track direction:

$$U_{\mathrm{eff}} = |\mathbf{U} \cdot \mathbf{n}_{\mathrm{trk}}| = U|\sin\alpha|$$

where n_{trk} is the horizontal unit normal to the satellite track and α is the angle between the ERA5 wind (equivalently, the inferred plume axis) and the local track tangent. This U_{eff} correction is applied only to the real DQ-1 emission calculations. When the track is perpendicular to the plume, U_{eff} = U; for an oblique crossing, the normal wind component is smaller than the full wind-speed magnitude and is the appropriate transport term for the cross-sectional flux calculation.

This real-case projection is separate from the 25° quality-control test. The 25° test compares the ERA5 wind direction with the vector from the stack to the fitted plume centre and rejects cases in which the observed enhancement is not plausibly downwind of the source; it does not require the track–wind angle to be close to 90°. For the real-overpass formal uncertainty, the wind contribution is evaluated as $u_{U_{eff}}/U_{eff}$. Uncertainty in wind direction, and hence in α, is not independently resolved by the present data and is retained as a structural transport uncertainty rather than being folded into a deceptively precise formal error bar.

In the original presentation, a wind speed and direction were printed above each map, but the conversion from an oblique track crossing to the transport velocity was not defined. The revision displays the ERA5 vectors, states the track orientation, and uses the track-normal component U_{eff} in every real-case flux calculation.

The projection is consequential rather than cosmetic. For the Mary overpass, the ERA5 wind magnitude is 4.15 m s−1, while the track-normal component used in the inversion is 2.14 m s−1 because the crossing angle is about 31°. Substituting the full magnitude into a flux estimate proportional to wind speed would make the wind factor about 1.94 times larger. For Anpara, by contrast, the wind magnitude is 7.90 m s−1 and U_{eff} is 7.81 m s−1 at an approximately 81° crossing, so the projection changes the transport term by only about 1 %. These real-data examples are shown through the revised Fig. 15 annotations and source data.

Equation (8) now propagates the reported formal terms associated with Â, σ̂, and U_{eff}. The resulting uncertainties are therefore conditional on the selected profile, background form, ERA5 field, and transport-height treatment. For example, Berezovskaya is reported as 314 ± 148 kg s−1 at N_Q = 15, and Luohuang as 1956 ± 791 kg s−1 at N_Q = 25. These intervals quantify the fitted/profile and effective-wind components included in Eq. (8); they are not relabelled as a complete truth uncertainty or as a validation against the inventory.

The real-case examples compare averaging scales within the same overpass because the same anchor and wind field are used across all six scales. They do not establish a complete-error-budget truth estimate. EPA hourly CEMS values and Climate TRACE monthly means remain post-selection external context; the revision no longer labels the EPA comparison as transport-time matched.

Changes in manuscript

Location in revised manuscript: Methods 2.2 and 2.5.3; Eq. (8); simulation setup in Sect. 3.1; Results 3.3.4

The revised text now distinguishes the prescribed wind speed U used in the controlled simulations from U_{eff} used in real DQ-1 emission calculations. The simulations assume a 270° wind perpendicular to the idealized track, so U_{eff} = U. For real overpasses, U_{eff} is the wind component normal to the satellite transect and is kept separate from the source-to-plume 25° wind-direction screen. These checks remove clear geometric mismatches but do not eliminate uncertainty in wind direction, plume rise, boundary-layer structure, or local background variability.

Major comment 5: Allan deviation analysis

5.1 Estimator, deviation, and white-noise reference

Reviewer comment

Allan deviation

Given the larger methodological problem above this seems almost minor, but I was not really convinced by the application of the Allan deviation as it was described in the paper. Typically such

plots are shown on a log-log scale, with the theoretical slope for white noise shown. This can be seen in e.g. Figs. 6 and 8 of (Amediek et al., 2017). (At very least, this τ —1/2 line should be shown.)

I was also a bit surprised by the formula given in Eq. 4, which is actually for the Allan variance, rather than its root, the Allan deviation. (There is also no explanation of what k and M are in the equation.) This formulation of the fixed-τ estimator of the Allan variance requires a new dataset for each value of τ . I think it might make more sense to use the overlapping Allan variance, which allows one to use all the data points in a given timeseries repeatedly. See e.g. Eq. 3 in Fix et al. (2018) for the both the formula and some discussion of its use.

Author response

This point was correct. We have replaced the original expression with the square root of an overlapping Allan estimator and, importantly, have made the averaging operation consistent with the actual signal-first processing. We follow the averaging-of-signals (AVS) scheme described for the MERLIN spaceborne IPDA lidar by Tellier et al. (2018, their Eq. 17): calibrated offline and online signals are averaged separately before the logarithmic DAOD retrieval is evaluated. Because every window contains the same N samples, the ratio of the two window means is identical to the ratio of the corresponding signal sums used in our implementation. The revised manuscript therefore writes the DQ-1 AVS operation explicitly before applying the overlapping Allan difference.

$$\begin{aligned} \hat{\delta}^{\mathrm{AVS}}_{N,j} &= \frac{1}{2}\ln\left(\frac{\langle q^{\mathrm{off}}\rangle_{N,j}}{\langle q^{\mathrm{on}}\rangle_{N,j}}\right) - \delta_{\mathrm{bias},N,j} \\ &= \frac{1}{2}\ln\left(\frac{\sum_{i=j}^{j+N-1} q_i^{\mathrm{off}}}{\sum_{i=j}^{j+N-1} q_i^{\mathrm{on}}}\right) - \delta_{\mathrm{bias},N,j}, \\ \mathrm{IWF}_{N,j} &= \frac{\sum_{i=j}^{j+N-1} q_i^{\mathrm{off}}\ \mathrm{IWF}_i}{\sum_{i=j}^{j+N-1} q_i^{\mathrm{off}}}, \\ X^{\mathrm{AVS}}_{N,j} &= 10^6 \frac{\hat{\delta}^{\mathrm{AVS}}_{N,j}}{\mathrm{IWF}_{N,j}}. \end{aligned} \tag{9a}$$

$$\sigma_A(N) = \left\{\frac{1}{2(n-2N+1)}\sum_{j=1}^{n-2N+1}\left[X^{\mathrm{AVS}}_{N,j+N} - X^{\mathrm{AVS}}_{N,j}\right]^2\right\}^{1/2} \tag{9b}$$

Here n is the number of native samples in one uninterrupted segment, N is the averaging count, i indexes samples within an AVS window, and j is the one-sample-shifted window origin. The angle brackets in Eq. (9a) denote the N-sample signal means; IWFN,j is weighted by the offline signal, consistent with the AVS treatment described by Tellier et al. (2018). Each Allan difference compares two adjacent non-overlapping N-sample AVS retrievals, while successive origins share observations; this is the overlapping construction advocated by Fix et al. (2018). Rejected samples and native-index gaps terminate a segment and are never bridged or concatenated.

The requested $N^{-1/2}$ reference is plotted explicitly. We use it only as the behaviour expected for stationary uncorrelated noise. Agreement with this slope is not taken as proof that atmospheric variability, surface effects, or retrieval correlations are absent. The main Fig. 2 curve is therefore described as AVS-based empirical scale-dependent retrieval variability, not as a pure instrument-noise calibration.

We also repeated the Allan calculation rather than simply relabelling the previous curves. The original desert plotting list contained 18 IDs, but 0000328283 duplicates 0000328032; the audit therefore contains 17 unique desert products. Sixteen desert products retain an uninterrupted segment longer than 1500 native samples after the full environmental, continuity, and facility screen. The forest display uses a deliberately different, exploratory cohort: five July–August products retained under the legacy retrieval-quality definition (finite XCO_2 and geolocation, clear-sky/land/day flags,

and SNR ≥ 7), with longest uninterrupted fragments of 152–195 samples. These five tracks were not treated as if they passed the full strict screen. Applying the land/water/snow, continuity, and facility exclusions leaves only one 154-sample forest segment, and no forest product meets the >1500-sample baseline criterion. We therefore retain the strict 16-track desert baseline as the only quantitative simulation input; the five-track summer forest panel is an exploratory legacy-quality scene comparison, and the single strictly screened forest segment is reported only in the screening audit.

For those 16 eligible desert tracks, the equal-product mean Allan deviation is 8.895, 6.347, 5.205, 4.501, 3.972, 3.587, 3.304, and 3.102 ppm for N = 5, 10, 15, 20, 25, 30, 35, and 40, respectively. Thus the mean decreases by about 65% between N = 5 and N = 40. Linear detrending leaves this strict long-segment baseline essentially unchanged: at N = 40 the mean is 3.10199 ppm instead of 3.10226 ppm, a change of only about 0.009%. This does not prove that the atmosphere is uniform; it only shows that a fitted linear trend is negligible for this particular long, screened desert cohort.

Compared with the original manuscript, the revised analysis distinguishes Allan deviation from variance, defines every index, applies AVS before differencing, and adds both the N−1/2 reference and the physical-support axis. The original 18-item list was also audited to 17 unique desert products, of which 16 meet the stated long-segment criterion.

The revised Allan analysis supports only one limited conclusion: averaging reduces empirical scale-dependent variability in the screened long desert sequences over N = 5–40. It does not independently select N_{det} or N_Q, and it is not used as a universal mission precision curve. Real power-plant detection uses a plume-excluded local Allan estimate from the same overpass, while the strict desert mean is used only as the conditional error input to the controlled sensitivity experiments.

At a fixed N, overlap is used along the window-position index. If $y_i^{(N)}$ is the AVS retrieval for the valid window beginning at index i, adjacent valid windows $y_i^{(N)}$ and $y_{i+1}^{(N)}$ contribute their squared difference. M is the number of valid adjacent pairs, and the Allan deviation is the square root of one half their mean squared difference. Screened discontinuities split a sequence, so no difference is formed across a gap.

The regional mean is an equal-product mean. Each eligible orbit product contributes one product-level Allan value at a given N before the across-product average is formed. A long product therefore does not dominate merely because it supplies more shifted windows. This aggregation choice, the unique-product audit, and the >1500-native-sample continuity rule are now stated together.

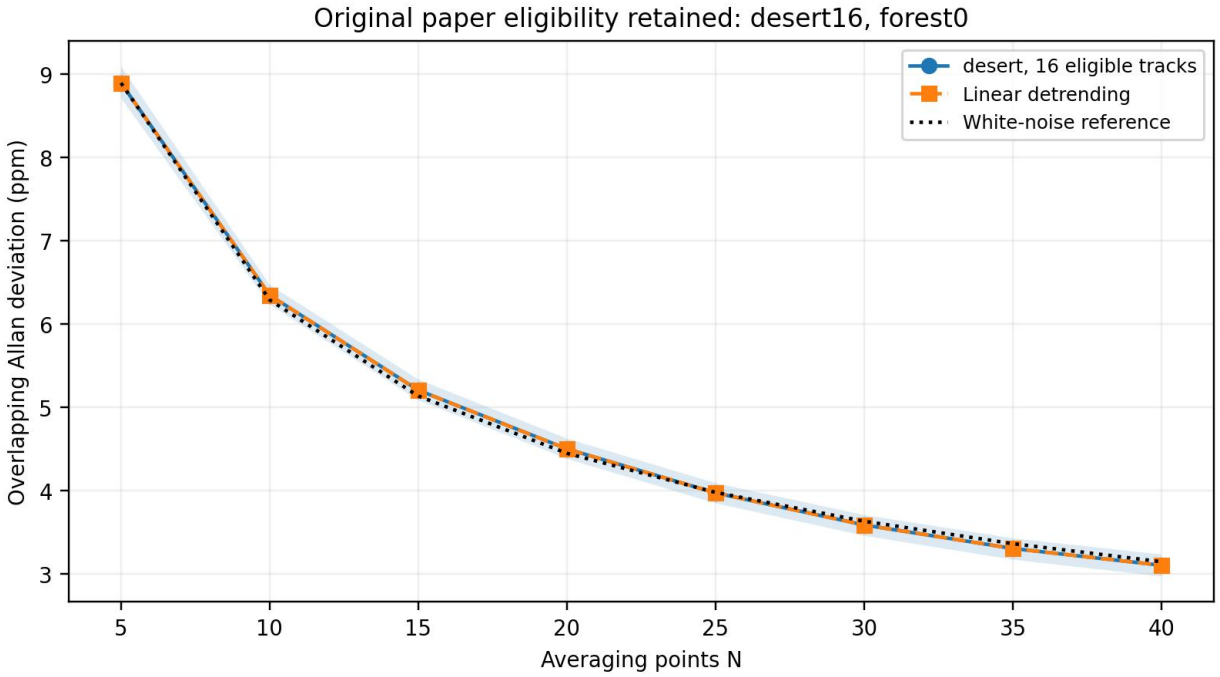

Figure R4. Strict baseline using the original manuscript continuity requirement. Sixteen screened desert products retain >1500 consecutive native samples; no screened forest product does. Raw and linearly detrended desert means are almost indistinguishable. The dotted $N^{-1/2}$ line is a white-noise reference, not a fitted instrument-noise model.

Changes in manuscript

Location in revised manuscript: Methods 2.3; Eqs. (9a)–(9b); Methods 2.5.1; revised Allan figure and caption; Results and Discussion passages that refer to the Allan-derived error input.

The revised text now defines the N-sample AVS retrieval directly from summed calibrated online/offline signals, then applies the overlapping Allan difference to adjacent AVS windows. All symbols (n, N, i, and j) are defined explicitly, the $N^{-1/2}$ reference is shown, and the screened 16-track desert curve remains the only regional quantitative input used in the controlled simulations. The forest curve is not treated as an equally supported precision calibration.

5.2 Surface heterogeneity, atmospheric gradients, and screening

Reviewer comment

Beyond these technical concerns, I was a bit confused about how this was applied. Is the assumption really that neither the albedo nor the true X_{CO_2} values change over single tracks for these regions? This seems very unrealistic. In the Arabian peninsula there is substantial oil and gas production, with associated emissions. Several of the tracks in the Boreal region go right over Lake Baikal, which would surely have an impact on the albedo, as would the presence of snow or the phenology of the larch forests in this region over the period from June through December. Even in the absence of local sources or sinks and albedo heterogeneity, the gradients of X_{CO_2} over zonal transects of over 1000 km can exceed 1 ppm simply due to synoptic-scale transport. The idea that these measurements should be uniform and represent only noise seems unreasonable.

Author response

We agree with the physical point in this comment. The revised manuscript no longer assumes that a whole Arabian-Peninsula or boreal track has constant albedo or constant true XCO_2, and we no longer describe all observed variability as instrument noise. The quantity is now called empirical

scale-dependent retrieval variability. It can include instrumental and retrieval noise, but also residual surface heterogeneity, transported CO_2, biological fluxes, and atmospheric gradients. The screening is intended to remove identifiable problems and test sensitivity to reasonable choices; it is not used to claim that the retained atmosphere is source-free or perfectly homogeneous.

We first re-audited every track that contributed to the original Allan figure. The audit uses finite XCO_2 and geolocation, the existing clear-sky/land/day and SNR ≥ 7 quality definition, ESA WorldCover 2021 as a static categorical land-cover reference, observation-day IMS snow and ice, a water-neighbourhood check, exclusion within three native samples of a land-cover transition, and a facility screen based on 2024 Climate TRACE facility-type CO_2 sources emitting at least 100,000 t yr^{-1}. Any rejected point or sampling discontinuity splits the sequence. WorldCover is not treated as a 2024 albedo measurement, and the source inventory is not treated as a complete map of all oil-and-gas activity.

The original broader forest audit confirmed the reviewer's concern about seasonal snow contamination: the 15 November product contained 4398 snow-classified points among 4441 quality-selected points (99.0 %), and the 22 November product contained 401 of 589 (68.1 %). These autumn/winter products are excluded. Revised Figs. 1b and 2b instead show five July–August products selected only under the legacy retrieval-quality definition for an exploratory comparison; their longest continuous fragments are 152–195 samples, and none meets the >1500-sample long-baseline criterion. This five-track display must not be confused with the strict forest sample: after the land/water/snow, continuity, and facility screens, only one 154-sample segment remains. The strict result cannot support a forest population curve, and neither forest set is used to define the simulation error model.

Response Table R1a. Audit of the 17 unique desert products in the original Allan plotting cohort. "Longest" is the longest uninterrupted segment after the 25 km substantial-facility screen; the >1500 column applies the original manuscript continuity requirement.

Product	Date	Quality pts	Snow pts	Water/neigh. pts	Longest pts	>1500
0000309494	2024-08-23	4136	0	23	2297	Yes
0000314308	2024-09-06	3549	0	54	1581	Yes
0000317297	2024-09-15	3852	0	35	1782	Yes
0000318471	2024-09-18	4884	0	33	2172	Yes
0000320349	2024-09-23	4667	0	14	1725	Yes
0000327311	2024-10-13	3738	0	80	1892	Yes
0000328032	2024-10-16	5406	0	32	2211	Yes
0000332766	2024-11-01	4400	0	25	1526	Yes
0000335278	2024-11-06	6083	0	26	926	No
0000336929	2024-11-10	4588	0	29	2828	Yes
0000337648	2024-11-11	5080	0	97	1838	Yes
0000339965	2024-11-17	4265	0	43	2718	Yes
0000346626	2024-12-06	4861	0	16	3531	Yes
0000348137	2024-12-10	3876	0	80	2486	Yes
0000349104	2024-12-13	5032	0	29	2822	Yes
0000352253	2024-12-22	4701	0	40	2795	Yes
0000352957	2024-12-24	4531	0	13	2490	Yes

Response Table R1b. The five July–August forest products shown in revised Figs. 1b and 2b under the legacy retrieval-quality definition. These are exploratory scene-comparison tracks, not the fully screened forest cohort. Their longest uninterrupted fragments are 152–195 native samples; none satisfies the >1500-sample quantitative-baseline criterion or supplies the simulation error input. Applying the full land/water/snow, continuity, and facility screen leaves a separate single 154-sample forest segment, as reported in Response Table R2.

Product	Date	Retained segments	Longest pts	Longest span (km)	>1500	Simulation input
0000290736	2024-07-01	1	152	51.46	No	No
0000292520	2024-07-06	5	195	66.00	No	No
0000292568	2024-07-06	4	176	59.53	No	No
0000308715	2024-08-20	2	153	51.76	No	No
0000312238	2024-08-31	2	176	59.58	No	No

The screening sensitivity itself is shown below rather than being left only in supplementary material. For the desert cohort, changing the substantial-facility exclusion radius from 10 to 25 to 50 km changes the N = 5 mean only from 9.0640 to 9.0199 to 8.9915 ppm, a total spread of 0.0725 ppm. At N = 40 the corresponding values are 3.1573, 3.1481, and 3.2136 ppm, a spread of 0.0655 ppm. Across all eight tested averaging counts from N = 5 to 40, the largest relative spread among the three facility-buffer curves is 2.1 %. Thus, the choice of 25 km changes the retained sample but does not create the overall decrease of Allan deviation with N. For the fully screened forest audit, only one 154-sample segment remains after the land/water/snow and continuity screen, and the same segment survives all three facility buffers. This single strict segment is distinct from the five-track legacy-quality forest display in Figs. 1b and 2b. A forest buffer-sensitivity curve is therefore not plotted; its sample count and endpoint values are retained in Response Table R2 for transparent accounting.

For the screened 25 km desert exploratory set, uncertainty in the scene mean is estimated by resampling whole tracks rather than treating overlapping Allan differences as independent replicates. The track-bootstrap 95% interval is 8.776–9.260 ppm at N = 5 and 3.031–3.265 ppm at N = 40. No corresponding population interval is reported for the screened forest result because only one original forest track survives the full screen. This is an important limitation rather than a missing error bar.

Response Table R2. Sensitivity of the original-track cohort to the successive screening choices. Each segment is evaluated first, segments are averaged within track, and tracks are then given equal weight. Changes in the forest rows also reflect changes in sample composition and are therefore not interpreted as a causal surface effect.

Scene	Screen	Tracks	Segments	N=5 (ppm)	N=40 (ppm)
Desert	Legacy quality	17	54	9.6101	3.3387
Desert	Land/snow/water screen	17	62	9.0533	3.1973
Desert	Facility buffer 10 km	17	63	9.0640	3.1573
Desert	Facility buffer 25 km	17	60	9.0199	3.1481
Desert	Facility buffer 50 km	17	51	8.9915	3.2136
Forest	Legacy quality	5	30	12.4387	4.2102
Forest	Land/snow/water screen	1	1	7.3924	4.1610

Forest	Facility buffer 10 km	1	1	7.3924	4.1610
Forest	Facility buffer 25 km	1	1	7.3924	4.1610
Forest	Facility buffer 50 km	1	1	7.3924	4.1610

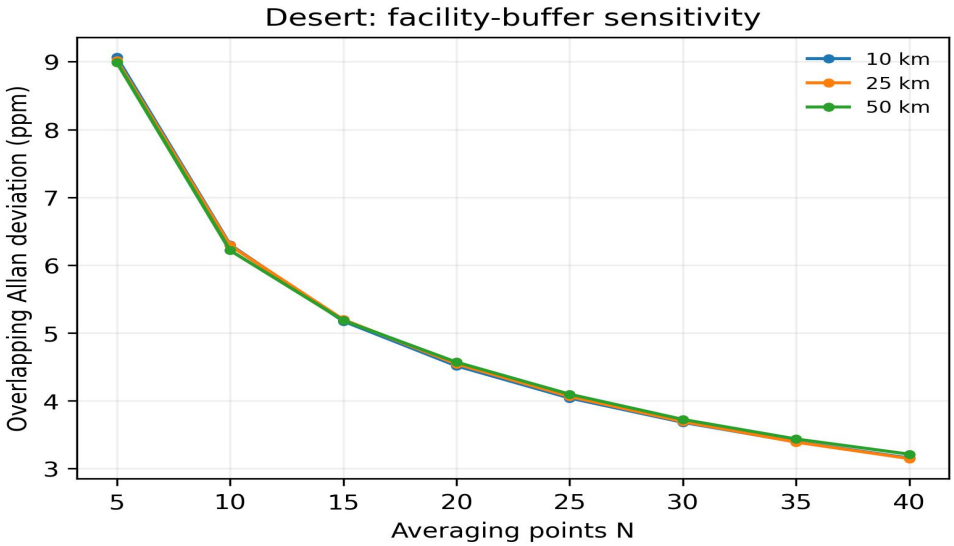

Figure R5. Desert equal-track mean facility-buffer sensitivity. The 10, 25, and 50 km curves differ by at most 2.1 % over N = 5–40. No forest panel is shown because all three buffers retain the same single 154-sample segment, which cannot support a cohort-level buffer-sensitivity estimate.

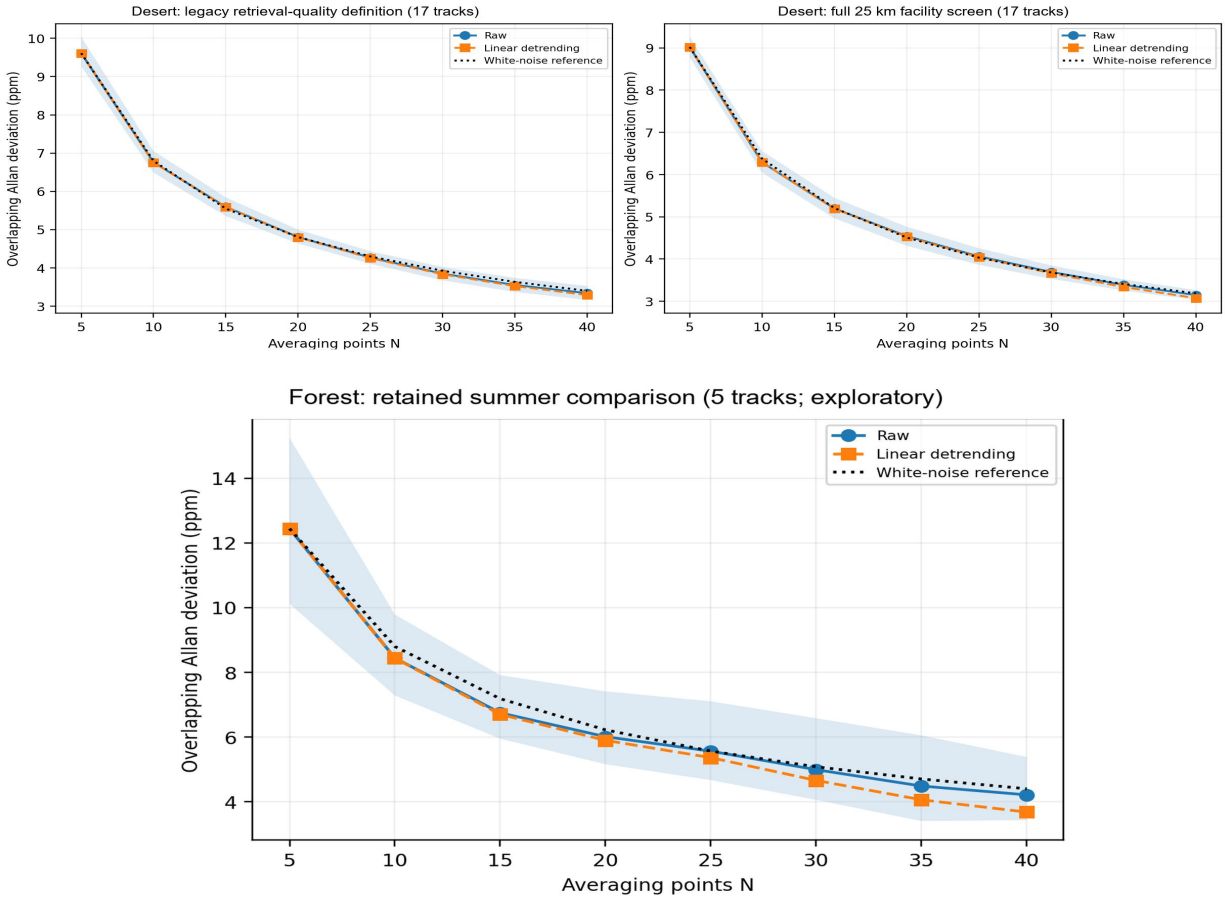

Figure R6. Raw-versus-linearly-detrended Allan deviation for the three cohorts with more than one contributing track: desert under the legacy retrieval-quality definition, desert after the full 25 km facility screen, and the five retained summer forest tracks evaluated under the legacy retrieval-quality definition. The shaded intervals are obtained by resampling whole tracks rather than overlapping Allan differences. The forest panel is exploratory and is not used as a regional precision baseline; the fully screened forest result is not plotted because only one short track survives.

The reviewer also raised synoptic-scale XCO_2 gradients. We tested this in two ways. First, raw and linearly detrended Allan calculations were made on exactly the same uninterrupted samples; screened gaps are never bridged and separate fragments are never concatenated into a regional transect. For the broad screened desert exploratory set (17 tracks, 60 local segments), the N = 40 mean changes from 3.148 to 3.075 ppm after detrending, a 2.33% decrease, although an individual short segment can change by as much as 36.9%. The much stricter 16-track >1500 baseline changes by only about 0.009% at N = 40. This contrast is why we do not use a single "gradient correction" as if it separated the atmosphere cleanly from measurement error.

Second, we held the total parent-interval coverage fixed while changing only the partition length. Sixty-one uninterrupted 200 km desert parent intervals were partitioned into four 50 km pieces, two 100 km pieces, or one 200 km piece, always using the same native indices and never joining across gaps. Thus each parent interval contributes the same total native-index coverage at every tested segment length. At N = 40, the raw means are 2.922, 3.005, and 3.026 ppm for 50, 100, and 200 km, respectively; after linear detrending they are 2.660, 2.976, and 3.022 ppm. The 200 km result is almost unchanged by detrending, whereas the 50 km result is more affected. There is no screened forest parent interval long enough for the matched 50/100/200 km test, so we report the comparison as unavailable rather than manufacture one by concatenating shorter pieces.

Comparison with the manuscript originally reviewed. The old text referred to Arabian desert and boreal forest tracks as homogeneous regions and attributed their variation mainly to random retrieval error. That language has been removed. The reviewer's examples proved important during the audit: two November forest products contained 99.0 % and 68.1 % snow-classified points after the original quality selection, and several forest transects crossed water or land-cover boundaries. These products cannot support a homogeneous-forest precision curve.

The revised manuscript separates two forest sets that answer different questions. Five July–August tracks, with uninterrupted fragments of 152–195 native samples, are retained only as an exploratory legacy-quality scene display. They do not meet the original >1500-sample continuity condition. After the full land/water/snow, transition, continuity, and facility screen, only one 154-sample forest segment remains. It is reported transparently in the audit and is not averaged with the five exploratory tracks or used to set the simulation error.

For the desert, the original list of 18 IDs contained one duplicate, leaving 17 unique products; 16 retain a >1500-sample uninterrupted segment after the full screen. Sensitivity to the substantial-facility buffer is small relative to the scale dependence: changing the radius from 10 to 25 to 50 km changes the N = 5 cohort mean from 9.0640 to 9.0199 to 8.9915 ppm, and the largest relative spread across N = 5–40 is 2.1 %. The matched-parent test also keeps identical native coverage while comparing 50, 100, and 200 km partitions, so its differences cannot be produced by selecting different pieces of the track.

We no longer claim a desert–forest instrument-precision contrast from comparable homogeneous populations, because such a forest population is not present after strict screening. The supported result is the screened desert scale dependence and its sensitivity bounds. The term used throughout the revised manuscript is empirical scale-dependent retrieval variability, explicitly allowing instrument, retrieval, surface, and atmospheric contributions. This change makes the simulation input more limited but removes an assumption that the data cannot verify.

The real-overpass validation is not affected by the absence of a forest population curve because it does not import either regional curve. For each plant, the local background is taken from the same track after excluding the plume region and detrending as specified. This local estimate enters D(N) and the background-cleanliness diagnostics. The regional audit and the plant-scale procedure are therefore evidence at different levels and are not pooled as interchangeable noise estimates.

The facility catalogue enters the regional audit only as a conservative mask for known substantial sources, not as an emission target. Changing the exclusion radius is therefore a screening-sensitivity test. It cannot guarantee that every smaller or unlisted source is absent, which is another reason the residual is called empirical retrieval variability rather than instrument noise.

Detrending is also treated as a sensitivity test. In the broad screened desert set, detrending changes the N = 40 mean from 3.148 to 3.075 ppm, whereas the strict long-segment baseline changes by about 0.009 %. Individual short segments can change much more. We therefore do not apply one regional trend correction as a universal separation of atmosphere and measurement error; local plant backgrounds are evaluated separately.

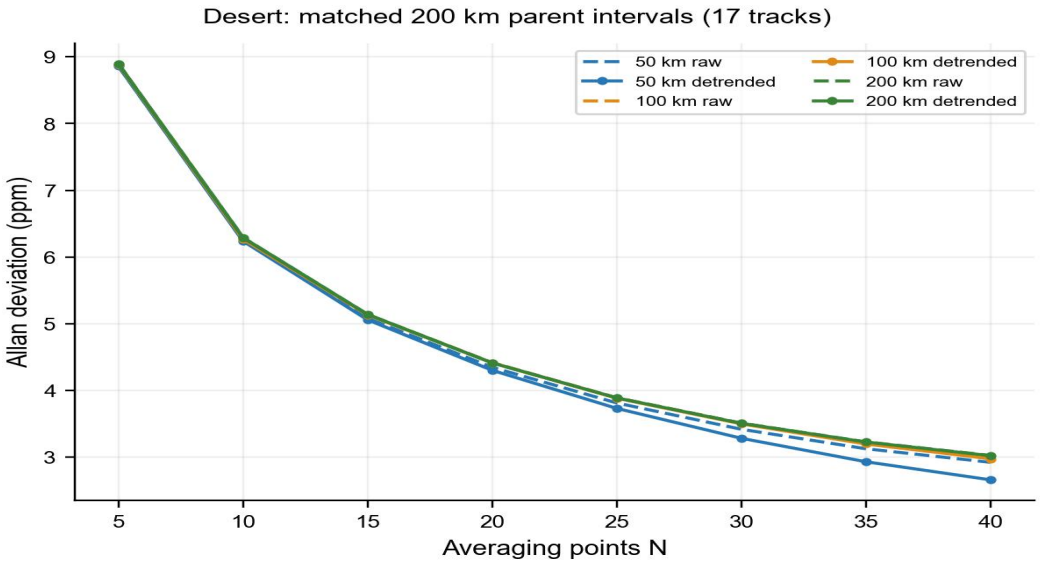

Figure R7. Matched-parent segment-length test for the screened desert cohort. The 50, 100, and 200 km estimates use the same 61 uninterrupted 200 km parent intervals across 17 tracks, and each parent contributes identical total native-index coverage at every segment length. No forest panel is shown because no screened continuous 200 km forest parent exists.

We also checked residual correlation rather than assuming that non-overlapping blocks are statistically independent. For the screened desert segments, the equal-track mean native-residual lag-1 autocorrelation is −0.012, with individual values from −0.177 to 0.140. The single surviving short forest segment gives 0.058, but its large-N block series is too sparse for a stable autocorrelation estimate. We therefore make no independence claim based solely on non-overlap; this diagnostic is reported numerically rather than as an additional Allan figure.

Because linear detrending can change an Allan estimate in a finite record, we also applied the same detrending operation to 1000 pure-white-noise realisations at 146, 292, and 584 native samples. At N = 40, detrending lowers the mean Allan deviation by 5.43%, 0.98%, and 0.10%, respectively. We therefore do not attribute every raw-versus-detrended difference to atmospheric transport, particularly for short records. The numerical control is retained in the response, but its separate non-observational diagnostic plot has been removed.

The albedo/SNR concern was evaluated separately, but the screened desert and forest samples do not provide sufficient matched return-signal support for a causal surface comparison. The native online/offline return fields are therefore used only as signal-magnitude diagnostics, are not equated with albedo, and are not used to calibrate or transfer the Allan curve. We removed the sparse return-regime and SNR-stratification plots from this response rather than presenting them as Allan validation.

Because the reviewer specifically mentioned Lake Baikal, snow, and seasonal larch conditions, we also performed an archive-wide screening check. Among 4577 June–December native products searched, eleven readable forest candidates contained more than 1500 consecutive retrieval-quality points before environmental screening, but none retained more than 1500 after the full land/water/snow, continuity, and facility screen. The July–August search yielded only short screened forest fragments of 122–195 samples. These fragments support provenance and scene comparison only; they are not used as a quantitative forest precision baseline.

Finally, the source screen has an important limitation that we now state explicitly. Facility centroids are useful for removing known large point sources, but basin-level oil-and-gas entries are not exact

wells or field boundaries, and no available inventory proves the absence of all oil-and-gas activity, biological fluxes, or advected CO_2 outside a chosen radius. The 10/25/50 km experiment is therefore a robustness test of the background selection, not a demonstration of a source-free atmosphere. This is also why we avoid calling the remaining Allan variability "instrument noise".

Changes in manuscript

Location in revised manuscript: Methods 2.3; revised Allan-data description and Fig. 2 caption; related limitations in Results, Discussion, and Conclusions. The response retains only direct Allan-deviation evidence: the screened Allan-deviation baseline with the $N^{-1/2}$ reference, the desert facility-buffer sensitivity, raw-versus-detrended curves for cohorts with replicate tracks, and the desert matched-parent segment-length comparison. Other diagnostics are summarized numerically rather than presented as additional Allan figures.

The revised manuscript no longer treats the boreal domain as a homogeneous calibration target. Revised Figs. 1b and 2b show five July–August tracks retained under the legacy retrieval-quality definition only as an exploratory scene comparison. The separate strict environmental/continuity/facility screen leaves one 154-sample forest segment, which is insufficient for a cohort-level precision estimate and is reported only in the audit. Neither forest result is used to set the simulation error; the quantitative input is restricted to the strict long-segment desert baseline.

5.3 How the Allan curve is used in the simulations

Reviewer comment

Even if we assume for a moment that the approach is reasonable: how were the values in Fig. 2 applied? Is it expected that the albedo difference between the desert and the Taiga makes no difference to the measurement precision? Or was the precision for only one of the regions (fit to a curve, perhaps) applied for all the modelling studies that followed? The concrete use of these results was not sufficiently clear.

Author response

The revised mapping is now explicit and uses only the screened desert baseline quantitatively. For a given averaging count N, the noiseless plume is sampled at the corrected 337.5 m native sample-centre spacing and averaged over the corresponding non-overlapping N-sample support. Only after this spatial averaging is a zero-mean Gaussian perturbation added with standard deviation σA(N) from the screened 16-track desert Allan curve. We do not apply a second 1/√N reduction, because σA(N) is already the empirical scale-dependent quantity assigned to that averaged observation. The forest results are not used to set the main simulation error.

This construction is deliberately described as a conditional effective-error model. Independent Gaussian errors are imposed for the controlled sensitivity calculation; independence and Gaussianity are not inferred from the overlapping Allan analysis. Real power-plant overpasses use local plume-excluded variability at the relevant scale instead of importing the regional desert curve. Thus the revised study does not assume that desert and forest measurement precision are identical.

To show why the Allan curve cannot by itself determine an emission-optimal averaging scale, we reran the reviewer-facing distance sensitivity with the corrected 337.5 m spacing. The setup is Q = 1000 kg s^{-1}, U = 2 m s^{-1}, stability class B, downwind distances of 1, 2.5, 5, and 10 km, N = 5-40, and 1000 noise realisations for each condition. As in the other controlled simulations, the wind direction is fixed at 270° and perpendicular to the idealized track, so U_{eff} = U and no real-case wind projection is applied. The fitted model contains the source amplitude plus a constant and a linear background term. The 16-track desert curve is the quantitative error input; the single 154-sample forest curve is shown only as an exploratory stress test.

The result is not a monotonic improvement with averaging. For the desert input, the N = 5 emission RMSE is approximately 88%, 105%, 114%, and 171% at 1, 2.5, 5, and 10 km, respectively, and increases as N is enlarged over the tested range. The exploratory one-track forest input shows the same qualitative behaviour, with N = 5 RMSE of about 73%, 87%, 95%, and 142%. The absolute RMSE values are specific to this deliberately difficult conditional setup and should not be read as a mission-wide performance estimate. The relevant point for the reviewer's question is the direction of the trade-off: although σA(N) decreases with N, the plume signal is simultaneously averaged over a wider physical support and the number of independent cross-plume samples falls. At 10 km, where the plume enhancement is weaker, this loss of spatial information is more severe.

The main paper therefore does not select a scale from the Allan curve alone. Local Allan variability and observed plume contrast define N_{det}. Quantification then retains the same plume anchor, reprocesses N = 5–30 with non-overlapping signal-first AVS blocks, and selects N_Q only among scales that pass the independent emission-quality gates.

The original description did not state whether the regional Allan curve was applied before or after spatial averaging or whether it was reduced again by 1/√N. The revised order is explicit: sample the noiseless plume, form non-overlapping N-sample averages, add one perturbation with standard deviation σA(N), and fit the emission. No second 1/√N reduction is applied.

Taken together, these results do not support interpreting the regional Allan minimum as an emission-optimal averaging scale. Allan variability contributes to the controlled error model and to local D(N), but final scales require plume information as well. This is why the revised paper reports a detection optimum, a separate qualification optimum, and an adjacent-scale stability label rather than one minimum-noise N.

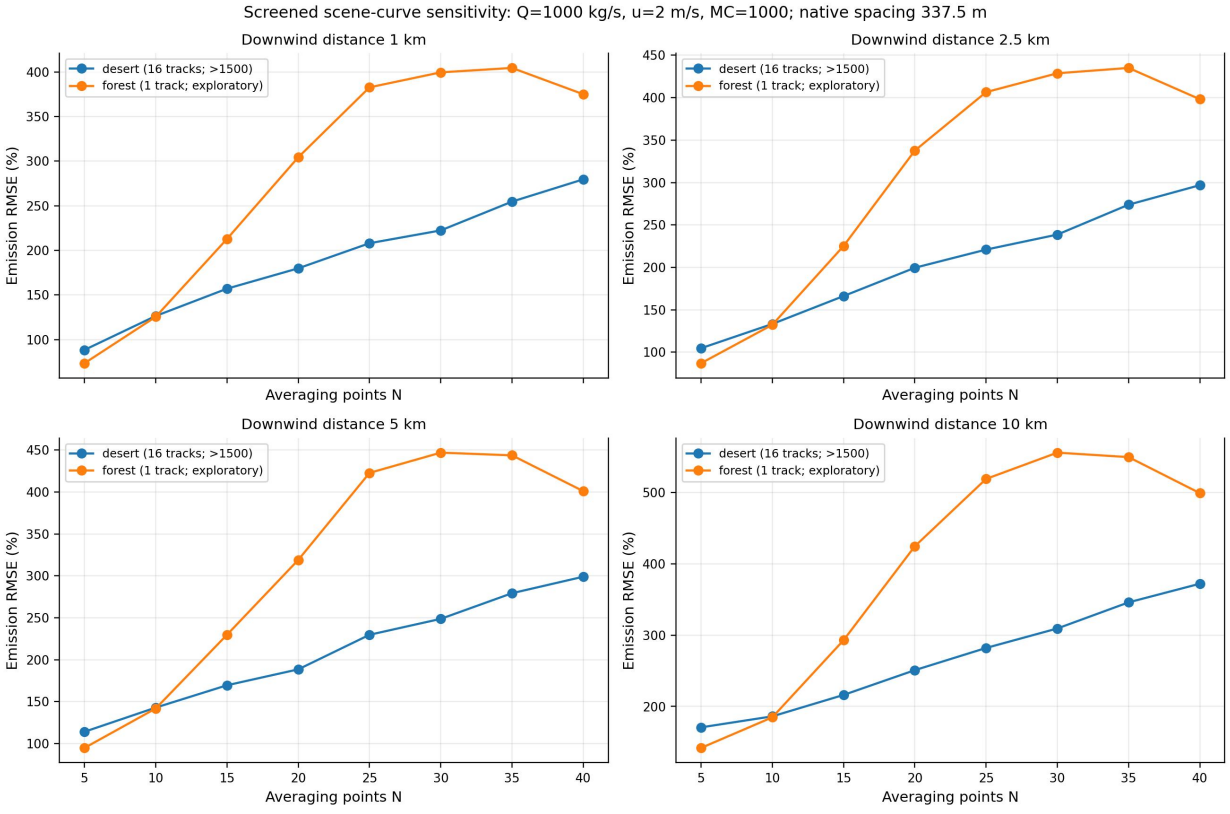

Figure R8. Point-source distance sensitivity rerun at the corrected 337.5 m native sample-centre spacing. Q = 1000 kg s⁻¹, U = 2 m s⁻¹, wind direction = 270° and perpendicular to the idealized track (therefore U_{eff} = U), stability B, N = 5-40, and 1000 noise realisations per condition. The

desert curve is the screened 16-track >1500 baseline. The forest curve is a single short-segment exploratory input and is not used to calibrate the main simulations. The reduction in Allan-derived variability with N is offset by plume smoothing and loss of spatial sampling.

Changes in manuscript

Location in revised manuscript: Methods 2.3-2.5; simulation-noise description; Results and Discussion passages on scale selection.

The manuscript now states the order of operations (native sampling → spatial averaging → scale-matched error addition → emission fit), identifies the desert Allan curve as the sole regional quantitative input, and separates this conditional error experiment from the observation-based local-background procedure.

Minor comment: References and data availability

Reviewer comment

Minor comments

I will refrain from writing out most of the minor notes I made while reviewing the study, given the major issues make it unacceptable for publication in its current form. The one thing I will note is the poorly formatted references: they were not in alphabetical order, they were missing DOIs, they (and all paragraphs) lacked indentation, they were not linked in the PDF (minor), at least one was missing (Amediek et al., 2017), and they did not always contain information consistent with how they were cited (see e.g. Wang et al., 2022; Zhang et al., 2024b; Reuter et al., 2019).

References

Amediek, A., Ehret, G., Fix, A., Wirth, M., Büdenbender, C., Quatrevalet, M., Kiemle, C., and Gerbig, C.: CHARM-F—a new airborne integrated-path differential-absorption lidar for carbon dioxide and methane observations: measurement performance and quantification of strong point source emissions, Appl. Opt., 56, 5182–5197, https://doi.org/10.1364/ao.56.005182, 2017.

Bovensmann, H., Buchwitz, M., Burrows, J. P., Reuter, M., Krings, T., Gerilowski, K., Schneising, O., Heymann, J., Tretner, A., and Erzinger, J.: A remote sensing technique for global monitoring of power plant CO_2 emissions from space and related applications, Atmos. Meas. Tech., 3, 781–811, https://doi.org/10.5194/amt-3-781-2010, 2010.

Fix, A., Quatrevalet, M., Amediek, A., and Wirth, M.: Energy calibration of integrated path differential absorption lidars, Appl. Opt., 57, 7501–7514, https://doi.org/10.1364/ao.57.007501, 2018.

Fuentes Andrade, B., Buchwitz, M., Reuter, M., Bovensmann, H., Richter, A., Boesch, H., and Burrows, J. P.: A method for estimating localized CO_2 emissions from co-located satellite XCO_2 and NO_2 images, Atmos. Meas. Tech., 17, 1145–1173, https://doi.org/10.5194/amt-17-1145-2024, 2024.

Han, G., Huang, Y., Shi, T., Zhang, H., Li, S., Zhang, H., Chen, W., Liu, J., and Gong, W.: Quantifying CO_2 emissions of power plants with Aerosols and Carbon Dioxide Lidar onboard DQ-1, Remote Sens. Environ., 313, 114368, https://doi.org/10.1016/j.rse.2024.114368, 2024.

Kiemle, C., Ehret, G., Amediek, A., Fix, A., Quatrevalet, M., and Wirth, M.: Potential of Spaceborne Lidar Measurements of Carbon Dioxide and Methane Emissions from Strong Point Sources, Remote Sensing, 9, 1137, https://doi.org/10.3390/rs9111137, 2017.

Kuhlmann, G., Broquet, G., Marshall, J., Clément, V., Löscher, A., Meijer, Y., and Brunner, D.: Detectability of CO_2 emission plumes of cities and power plants with the Copernicus

Anthropogenic CO_2 Monitoring (CO2M) mission, Atmos. Meas. Tech., 12, 6695–6719, https://doi.org/10.5194/amt-12-6695-2019, 2019.

Reuter, M., Buchwitz, M., Schneising, O., Krautwurst, S., O'Dell, C. W., Richter, A., Bovensmann, H., and Burrows, J. P.: Towards monitoring localized CO_2 emissions from space: co-located regional CO_2 and NO_2 enhancements observed by the OCO-2 and S5P satellites, Atmos. Chem. Phys., 19, 9371–9383, https://doi.org/10.5194/acp-19-9371-2019, 2019.

Shi, T., Han, G., Ma, X., Pei, Z., Chen, W., Liu, J., Zhang, X., Li, S., and Gong, W.: Quantifying strong point sources emissions of CO_2 using spaceborne LiDAR: Method development and potential analysis, Energy Convers. Manage., 292, 117346, https://doi.org/10.1016/j.enconman.2023.117346, 2023b.

Wang, N., Zhang, K., Shen, X., Wang, Y., Li, J., Li, C., Mao, J., Malinka, A., Zhao, C., Russell, L. M., Guo, J., Gross, S., Liu, C., Yang, J., Chen, F., Wu, L., Chen, S., Ke, J., Xiao, D., Zhou, Y., Fang, J., and Liu, D.: Dual-field-of-view high-spectral-resolution lidar: Simultaneous profiling of aerosol and water cloud to study aerosol–cloud interaction, Proc. Natl. Acad. Sci. USA, 119, e2110756119, https://doi.org/10.1073/pnas.2110756119, 2022.

Wolff, S., Ehret, G., Kiemle, C., Amediek, A., Quatrevalet, M., Wirth, M., and Fix, A.: Determination of the emission rates of CO_2 point sources with airborne lidar, Atmos. Meas. Tech., 14, 2717–2736, https://doi.org/10.5194/amt-14-2717-2021, 2021.

Zhang, H., Han, G., Chen, W., Pei, Z., Liu, B., Liu, J., Zhang, T., Li, S., and Gong, W.: Validation method for spaceborne IPDA lidar XCO_2 products via TCCON, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., 17, 16984–16992, https://doi.org/10.1109/jstars.2024.3418028, 2024a.

Zhang, X., Yang, H., Bu, L., Fan, Z., Xiao, W., Chen, B., Zhang, L., Liu, S., Wang, Z., Liu, J., Chen, W., and Lee, X.: Estimation of diurnal emissions of CO_2 from thermal power plants using spaceborne integrated path differential absorption (IPDA) lidar, EGUsphere [preprint], https://doi.org/10.5194/egusphere-2024-3152, 2024b.

Zhang, X., Yang, H., Bu, L., Fan, Z., Xiao, W., Chen, B., Zhang, L., Liu, S., Wang, Z., Liu, J., Chen, W., and Lee, X.: Estimation of diurnal emissions of CO_2 from thermal power plants using spaceborne integrated path differential absorption (IPDA) lidar, Atmos. Chem. Phys., 25, 6725–6740, https://doi.org/10.5194/acp-25-6725-2025, 2025.

Author response

The list is now alphabetized, DOI links have been added where available, malformed or unused entries have been corrected or removed, and Amediek et al. (2017) is included. The roles of Wang et al. (2022b), Zhang et al. (2024a, 2025), Reuter et al. (2019), and the primary Gaussian-plume and Allan references were checked against the revised text. Satellite-image panels also carry the relevant Esri or NASA GIBS/EOSDIS credit.

To support independent inspection, the revised submission provides one representative original Level-2C product corresponding to the Anpara example in Fig. 15 and a reviewer-access package of derived data. The separate AMT supplement now reports all 38 eligible flux cases in Table S1, including the processed case identifier, matched H5 UTC observation date, N_{det}, N_Q, $\hat{Q} \pm u_Q$, fit metrics, U_{eff}, downwind distance, background excursions, adjacent-scale stability, and the exact external comparison period. Table S2 reports the N_Q decision-rule sensitivity and Table S3 reports the dispersion coefficients. Table 3 is explicitly labelled as an illustrative 20-case subset rather than the basis for population frequencies.

Changes in manuscript

Location in revised manuscript: References; Data availability

One representative original DQ-1/ACDL Level-2C orbit product corresponding to the Anpara example in Fig. 15 and the non-restricted derived source data underlying the reported figures and tables are available in the persistent Zenodo record at https://doi.org/10.5281/zenodo.23874343. This concept DOI represents all versions and resolves to the latest version of the record. The derived materials include the plant/overpass index, scale-specific Allan and fit diagnostics, controlled-model width-scaling results, fixed-anchor block-AVS profile data, six-scale flux-quality and background-cleanliness results, cluster-robustness summaries, ERA5 wind fields used in Fig. 15, and the post-selection external-comparison fields used in Table 3. The wider DQ-1/ACDL Level-2C archive remains subject to the mission data-use and access policy.