

Quantifying methane emissions by combining multi-instrument airborne observations with regional atmospheric transport modelling: application to Madrid waste sites

Review Comment

Summary

The manuscript revisits the CoMet 2.0 Madrid waste-site dataset previously analysed by Krautwurst et al. (2025) with the cross-sectional flux method, and replaces that approach with a χ^2 -based inversion that fits a high-resolution WRF-GHG simulation to three independent CH₄ measurements (CHARM-F lidar XCH₄, MAMAP2DL column anomalies, and JIG in situ mixing ratios). Each waste site is represented by its own scalable tracer, together with a CAMS-based background and fixed EDGAR emissions elsewhere in the domain; a time-shift parameter compensates for a timing mismatch between the modelled and observed rotation of the wind field that occurred shortly before the overflights. Emission rates of $4.0 \pm 1.2 \text{ t h}^{-1}$ (Pinto) and $4.7 \pm 0.7 \text{ t h}^{-1}$ (Valdemingómez) are obtained, with the dominant uncertainty attributed to transport-model choice (estimated from the spread across five PBL schemes) rather than to measurement noise. These values are appreciably lower than, but consistent with, an accumulation-corrected re-analysis of the earlier cross-sectional estimate.

Recommendation

I find the approach timely and well motivated: explicit transport modelling is a sensible way to handle the non-steady, rotating wind field that undermines the cross-sectional method here, and the explicit comparison against the revised Krautwurst et al. estimate is a nice demonstration of added value over the existing literature. Beyond that, I'd note that multi-instrument comparison is itself a critical open question for the field right now: satellite observations are becoming the primary tool for GHG point-source monitoring at scale, and knowing how well independent instruments and methods agree — and where and why they diverge — is essential to trusting and correctly interpreting satellite-

derived emission estimates. Work that brings lidar, imaging-spectrometer, and in situ data together within a common transport-model framework, as this manuscript does, therefore has value beyond the two Madrid sites themselves, as a contribution to that broader validation effort.

That said, addressing the points below — in particular the instrument-weighting question and the two structural uncertainty concerns — will require additional analysis and diagnostics rather than wording changes alone, so I recommend major revisions. I want to be clear that this reflects the scope of the requested analysis, not a reservation about the paper's underlying merit: it's a strong, well-executed study, and I expect it to be a good paper once these points are worked through.

Major Comments

Here's the updated version, with that idea folded into the discussion:

Major comment

My main concern is that the paper doesn't really substantiate the "multi-instrument" framing in its own results. The single-instrument fits are relegated to Appendix D (Tables D1–D3), and the combined total emission rate sits in an interesting position relative to them that deserves discussion in the main text:

	Total [t h^{-1}]	Uncert. [%]
CHARM-F only	9.5	17
MAMAP2DL only	7.7	18
JIG only	15.7	4
Combined	8.7	18

The combined estimate sits close to the two remote-sensing-only estimates, while JIG alone very nearly doubles the total emission rate — yet reports by far the tightest uncertainty of the four. This is counterintuitive on its face: JIG contributes the fewest data points of the three instruments and, being an in situ point sensor along a single flight track, only has local sensitivity to the plume at one altitude and location, rather than the more complete spatial/column coverage

the two remote-sensing instruments provide. That JIG alone nonetheless returns the highest precision suggests its "4%" is measuring something narrower than the other rows — likely just the model's sensitivity to PBL-scheme choice along the (few) points where the aircraft happened to sample, rather than a real constraint on the total source strength. I'd like the authors to comment directly on this:

- What relative weight does each instrument effectively carry in the single combined χ^2 ? This is presumably set by the assumed measurement uncertainty per data point together with the number of points contributed by each instrument, but neither is stated. Please add these explicitly, ideally in Table 1 or 2.
- Is in situ point sampling simply less well suited to constraining an integrated source-strength estimate for a distributed pair of sources than column/remote-sensing data are, given its limited spatial sampling and lack of column integration? If so, is that reflected anywhere in how JIG is weighted in the combined fit, or does the χ^2 framework treat it as equally informative per data point regardless of this limited spatial sensitivity?
- More generally, does the PBL-scheme spread used in Sect. 2.6 adequately capture the uncertainty for an instrument with such limited spatial coverage, or does it systematically understate it? I'd actually expect the opposite bias to the one implied by the tight 4%: since JIG samples the smallest atmospheric volume of the three instruments, it should arguably be the most sensitive to local transport and turbulence errors in the model, not the least. A column-integrated or swath-averaged quantity (CHARM-F, MAMAP2DL) partially averages out small-scale transport noise that a point in situ measurement does not.

Two further, structural concerns with the uncertainty budget:

- Shared bias across the ensemble. Appendix B shows all five WRF runs share the same offset against the BAHAMAS wind measurements (~ 1 m/s, $\sim 10^\circ$). Since this bias is common to every ensemble member, the PBL-scheme spread (Sect. 2.6) cannot detect it — the quoted 18% reflects sensitivity to scheme choice, not a full transport-error budget, and should be described as such.

- Time-shift/flux degeneracy. The best-fit Δt (≈ -53 min) spans more than half the remote-sensing window. Sampling the tracer field earlier or later in its build-up changes the modelled mass independent of true source strength, so Δt and (S_{Pinto} , S_{Valde}) are natural candidates for correlation. The single-parameter uncertainties from the χ^2 curvature assume separability; I'd like to see a correlation matrix or a Δt - S contour before treating those uncertainties as final.

Finally, one concrete suggestion: applying the cross-sectional flux method directly to the WRF best-fit simulated enhancement field, rather than only to the real observations, would be a cheap and compelling validation. If that exercise reproduces the same over-estimate seen in Krautwurst et al.'s cross-sectional analysis of the real data, it directly confirms — using the authors' own simulation as ground truth — that the accumulation/rotation effect, rather than some other difference between analyses, explains the discrepancy between the two methods.

Bringing Table D1–D3 (or a condensed version, as above) into the main results section, alongside a paragraph addressing these points, would in my view be the single change that most improves the paper — and would turn an interesting but buried appendix result into a genuine methodological contribution about combining instruments of very different character.

Minor Comments

1. Introduction: The paper should situate itself more in previous airborne campaigns targeting emissions using inversion frameworks. Most notable is the work by CarbonMapper, e.g.
 1. Cusworth et al, 2022: Quantifying methane emissions from United States landfills
 2. Cusworth et al, 2020: Using remote sensing to detect, validate, and quantify methane emissions from California solid waste operations
2. I noticed that the paper does not cite any references for HYSPLIT and Sentinel-5p. Make sure to cite appropriate resources.
3. Line 100: You also assume constant emissions for the duration of the simulation.

4. Line 116: Great job on the video supplement!
5. Table 5: Be more expressive with your table headings. The definitions given here are very technical. I would move the PBL details to the SI and explain physical differences of the schemes in the main text, such that the reader understand the extend of the uncertainty provided.
6. L246: State if the 9.6 is a combined

Technical Comments

A copyedit pass would help. One recurring pattern: the word "very" appears four times in the manuscript, and in each case it's a filler that doesn't add meaning — either cut it or replace it with something more precise. Here are a few examples of colloquial phrasing I noticed:

- Introduction: "offers very promising perspectives to quantify greenhouse gas emissions" → "offers promising perspectives..." (or be specific about what makes them promising).
- Sect. 2.4: "...are independent of each other and use very different observational techniques." → "...use different observational techniques" (they're independent measurement principles regardless of degree).
- Sect. 3 / Fig. 3 discussion: "...one or two ground traces occasionally were very close to the sources." → give an actual distance if available, or just "close to the sources."
- Conclusions: "Nonetheless, very similar results are obtained when removing those measurements..." → "similar results are obtained..." — or, better, state the actual emission rate from that sensitivity test so the reader can judge "similar" for themselves.
- L316: "This pulls up the fit, resulting in a higher flux estimate." → e.g. "This biases the fit toward higher values, resulting in a higher flux estimate."
- L314: "...the Valdemingómez scaling factor has to compensate..." → "...must compensate..."

- L299 (Appendix B): "This comparison is, of course, only a snapshot..." → drop "of course."
- L47: "we are convinced that the efficient combination..." → a more measured "we argue" or "we contend."
- L94: "we tried using several tracers" → "we tested the use of several tracers" or "we initially employed several tracers."