RC2 Referee comment on egusphere-2026-5235 (Meier et al.)

„Evaluating metre-scale methane plume simulations with MicroHH against AVIRIS-4 remote sensing observations”

MAIN COMMENTS

M1. Comparison of the modelled vertical columns with observations along slant light paths (Sects. 2.5, 3.3-3.7, Fig. 8)

Section 3.7 shows that the modelled VCDs are not directly comparable to the AVIRIS-4 columns and that the GPCDs reproduce the double plumes, R^2, widths and maxima much better. However, the comparisons in Sects. 3.3-3.6 (Figs. 4-7) are based on the VCDs, and the differences are attributed to parallax (L. 318-323, 359-360) without verifying this using the GPCDs. The GPCD analysis is based on a single overpass (18 September 2024, 13:33 UTC). I understand the argument that, with slant light paths, the result is sensitive to the orientation of the plume and that the modelled and observed plume directions differed during the other overpasses (L. 264-270). Since 135 snapshots are available per release, the authors could select, for each overpass, those snapshots whose plume direction matches the observation, or rotate the modelled field around the source. The GPCD could then be applied to all overpasses in this manner (at least for the Gaussian fit statistics and the PSD). Alternatively, the authors could explain why this is not necessary or not possible.

In either case, the limitation that the VCD-based comparison is not fully like-for-like should be stated clearly, and the conclusion that MicroHH reproduces the observed plume morphology should be worded accordingly.

M2. Wind forcing and its consequences for plume variability (L. 153-156, L. 198-201, L. 259-266, L. 301-305, L. 350-351, L. 469-470, Fig. 3)

(a) Wind forcing and plume variability. The differences in plume morphology (R^2, width, meandering, flux variability) are attributed to parallax, sample size and surface effects, but the wind forcing is not discussed. The model is nudged towards a 15-min smoothed lidar profile, and the wind direction measured by the anemometer at 1.5 m is applied to the entire profile (L. 153-156). As an alternative, the lidar direction could be corrected for the 180-degree ambiguity using the anemometer (e.g. Knoop et al., 2021), which would retain the vertical profile of the wind direction. This probably suppresses variability on time scales of minutes to hours, including changes of wind direction with time and height (directional shear). Such variability could make real plumes more irregular and could contribute to double plumes, which are suppressed in the simulations by construction. This is consistent with the more irregular lateral structure of the observed

plumes (L. 350-351) and with the diverging modelled and observed plume directions during most overpasses (L. 264-266). Please discuss to what extent the forcing may explain the lower variability of the modelled plumes, whether wind shear may contribute to the observed double plumes, and what this means for the claim that the model reproduces realistic plume morphology. The statement at L. 469-470 should specify which time scales are represented, because a smoothed forcing contains only the slow variations. Finally, the reason for using the anemometer (direction ambiguity of the lidar when averaging to 20 s) is only given in Sect. 2.3. Please mention it already in Sect. 2.1.

(b) Independence of the temporal TKE comparison. The model is nudged towards a smoothed profile of the same wind lidar that is used for the temporal TKE spectra in Fig. 3 (L. 198-201, L. 301-305). The agreement at the lowest frequencies (periods longer than about 15 min) is therefore partly imposed by the forcing and is not an independent validation. Please state which part of the spectrum is independent of the forcing.

M3. Wind speed and along-plume flux (L. 244-246, L. 336-338, L. 375-377, Figs. 1, 2, 5, A2 and the supplement)

(a) Observed wind speed. It is not described how the wind speed of the observations is determined, neither for the flux calculation (L. 244-246 only describes the model) nor for the "Wind: x +- y m/s" labels in Fig. 2 and in the supplement (data source, height, averaging time). In the supplement (e.g. Fig. S3), the observed and modelled wind speeds differ considerably (2.8 vs. 1.6 m/s), and the sign of the difference changes between overpasses (e.g. Figs. S1, S3, S5 and Fig. 2). Please also explain how the uncertainties are determined. The modelled values of +- 0.0 m/s seem unrealistically small, and it is unclear whether the modelled and observed uncertainties (e.g. +- 0.5 m/s) are comparable quantities.

(b) Height of the wind speed. It is unclear at which height the wind speed for the flux is taken. The text states the "prescribed source height" (L. 245) and that plume rise is neglected (L. 338), whereas the source height is prescribed as the release height plus the calculated plume rise (L. 375-377, Table 1). If only the release height is used, the wind speed is probably too low, because wind speed usually increases with height. Meier et al. (2026, AMT) found that source-height winds work well for small releases but underestimate the flux for large ones. Please clarify which height is used and consider the wind at the plume-rise-corrected height (or averaged over the plume layer), or discuss the effect on the flux offsets.

(c) Low wind speed. The wind speed is only 1.4 m/s in Fig. 1 and 1.6 m/s in Fig. 2. Varon et al. (2018) restrict the cross-sectional flux method to U10 >= 2 m/s. This limit refers to U10 and cannot be applied directly here. Nevertheless, please state down to which wind speed the flux calculation is considered valid and justify its use at 1.4-1.6 m/s.

(d) Accuracy of the flux. The along-plume flux is used to compare plume intermittency, but its accuracy is not evaluated. Since the release rates are known, the flux averaged over all cross-sections should reproduce them. Please show the true release rate in Figs. 5 and A2 and report the release rate determined from the mean flux over all cross-sections (with its spread) for both the model and the observations. In Fig. 5, the modelled flux appears closer to the release rate than the observed flux (except for the 53 kg/h release). Please discuss why (e.g. wind speed, retrieval bias, R^2 filter). Please also comment on the difference between the estimated emission of 205 +- 20.5 kg/h in Fig. 1 and the release rate of 290 kg/h. For the model, such a comparison tests the flux method itself (wind speed choice, neglected plume rise). For the observations, it links the results to the emission estimates of Meier et al. (2026, AMT).

M4. Sample-size matching (L. 324-329, Figs. 4, 5, A1, A2, L. 551)

The matching uses one random draw without replacement at each along-plume position. The result therefore depends on the draw, and most of the 135 snapshots are discarded. I suggest keeping the full-sample statistics as the main result and repeating the random draw many times. Please show where the observed values lie within the distribution of the matched model draws (e.g. as a percentile). This would support the statements that the variability "became much more comparable" and that the results "do not provide robust evidence that MicroHH systematically underestimates plume intermittency" (L. 328-329), as well as the statement that the lack of intermittency is "largely a sampling effect" (L. 324 and L. 551). Please also show the number of valid cross-sections per along-plume position in Figs. 4, 5, A1 and A2 (e.g. as an additional small panel, or at least a table in the supplement).

M5. Strength of the conclusions (Abstract L. 8-10, L. 546-557)

The conclusions are formulated more strongly than the results support. Plume width is listed as reproduced (L. 546-548, also in the abstract L. 8-10), but Sect. 3.3 states that the widths agree only close to the source and that the observed plumes are generally wider (L. 313). The maxima agree well only for the 57 and 53 kg/h releases (L. 314-315). The restriction in the abstract ("most releases", "when inputs are well constrained and observational geometry is considered") is missing in the conclusions. The list of reasons for the remaining differences (L. 552-554) also omits the wind forcing (comment M2) and the viewing geometry (comment M1). Finally, only four releases at one site under low wind speeds (1.4-2.8 m/s) are evaluated. Higher wind speeds, other atmospheric stability and area sources (see Bue et al., 2026, L. 435) are not tested. Please restrict the claim that MicroHH can generate realistic synthetic plume imagery to the conditions tested, or discuss this explicitly in Sect. 4.1.

SPECIFIC COMMENTS

S1. L. 1-3, L. 14 and L. 29 onwards (motivation)

The motivation for this study is stated only in passing in the abstract, in the last sentences (L. 14: "Overall, MicroHH can generate realistic high-resolution synthetic plume imagery for retrieval evaluation and machine-learning applications [...]"). In the introduction (L. 29 onwards), however, it is a central argument: Synthetic plumes are increasingly used to train ML methods, but their realism is rarely validated against observations. I fully agree with this argument and consider it an important reason why this study is relevant. I therefore suggest moving it to the beginning of the abstract, directly after the first sentence. The first sentence ("[...] creating opportunities for emission quantification but requiring models that reproduce turbulent plume morphology at comparable resolution") does not make clear why such models are required. Stating the problem early (synthetic plumes are used for retrieval evaluation and ML training, but their realism is not validated) would make the aim of the paper clear immediately.

S2. L. 20-23 (references)

I suggest adding examples. For satellites, these could be Sentinel-2 (Varon et al., 2021; https://doi.org/10.5194/amt-14-2771-2021), PRISMA (Guanter et al., 2021; https://doi.org/10.1016/j.rse.2021.112671), EnMAP (Roger et al., 2024; https://doi.org/10.1109/TGRS.2024.3352403), GHGSat (Jervis et al., 2021; https://doi.org/10.5194/amt-14-2127-2021) and WorldView-3 (Sánchez-García et al., 2022; https://doi.org/10.5194/amt-15-1657-2022). For airborne systems, examples are AVIRIS-NG, MethaneAIR and MAMAP2D-Light (Krautwurst et al., 2025; https://doi.org/10.5194/acp-25-14669-2025). I name these from the top of my head. There may be newer or more suitable references, and the authors do not need to cite all of them.

S3. L. 25-27, L. 66-81 and L. 84-86 (introduction, optional)

The point that fine-scale structures are shaped by turbulent mixing, wind shear, surface heterogeneity and obstacles is made twice in the introduction (L. 25-27 and L. 84-86). The introduction is also rather long, and the list of previous MicroHH applications (L. 71-81) could be shortened. Both points are at the authors' discretion.

S4. L. 105-106 (wind lidar)

No technical description is given for the ZX 300 Doppler wind lidar. Please state the measurement principle (e.g. scanning strategy and probe volume), the measurement heights and the typical uncertainty. This matters because the lidar data drive the model and enter the temporal TKE spectra. A conically scanning lidar averages over a probe volume that grows with height and may damp turbulent fluctuations. Please discuss the consequences for the comparison with a single model grid point.

S5. L. 111-112 (emission estimation)

The authors refer to Meier et al. (2026, AMT) for the data processing and emission estimation. In this manuscript, an along-plume flux is computed from Gaussian fits and a wind speed (Sects. 2.4.3 and 3.4), which resembles the cross-sectional flux method. Please state explicitly which method and which wind speed definition are used here, and whether they are identical to those in Meier et al. (2026), where both the cross-sectional flux and the integrated mass enhancement method were applied depending on the plume.

S6. L. 132-134 (domain height)

Please confirm that the plume and the modelled boundary layer remained well below the model top in all simulations and add a short justification why a vertical domain height of 2038 m is sufficient. Figure C2 shows that the ERA5 boundary-layer height stays below about 1100 m, and the analysed part of the plume (first 50 m from the source) is probably far below the model top. This makes the choice plausible. However, Figure C2 is only given in the appendix and is based on reanalysis data. Since the authors found the ERA5 winds to be biased at this site (Sect. 2.3), the reliability of the ERA5 boundary-layer height is also unclear. Optionally, you could give the modelled boundary-layer height or the maximum plume height.

S7. L. 209-213 (PSD of the observed plumes with missing values)

Missing values in the AVIRIS-4 images are replaced with zero before the PSD is calculated. It is not clear whether this is done before or after the mean is removed, how large the fraction of missing pixels is in the 50 m x 75 m plume windows, and whether the modelled plumes were masked in the same way. Zero-filling can introduce artificial steps and reduces the variance, which affects the PSD of the observations but not that of the model, and thus the comparison in Fig. 7. Please state the order of the processing steps and the fraction of missing pixels, and show that the result is robust (e.g. by excluding windows with many missing pixels or by applying the same mask to the model).

S8. L. 237-246 (Gaussian fit and R^2 filter)

Gaussian fits are applied to cross-sections that visibly deviate from a Gaussian shape (e.g. the double structure at 45-51 m in Fig. 1b). A lower R^2 for double-peaked cross-sections is expected by construction (L. 318-323), so R^2, width and maximum in Fig. 4 partly reflect the parallax geometry rather than the plume morphology. I understand that the Gaussian fit is used because it provides width, shape and displacement of each cross-section. For the flux, however, a numerical integration of the enhancement above the background (Riemann sum) could be calculated in addition. It does not require a Gaussian shape and should conserve the integrated mass also for double-peaked cross-sections, as long as both branches are inside the integration window. It would also allow keeping cross-sections with R^2 < 0.5 for the flux statistics and thus increase the number of samples per along-plume position. Please justify the threshold of R^2 >= 0.5, which seems low to me, and report how many cross-sections are discarded for the model and for the observations. As the modelled cross-sections are more Gaussian (Fig. 4), the filter may retain different fractions of cross-sections and thereby affect the comparison of width and flux statistics.

S9. Fig. 3 and L. 297-305 (inertial subrange)

The grey area labelled "inertial subrange" in Fig. 3 is not defined. The text gives no wavenumber limits and no criterion for choosing them (e.g. a theoretical argument based on the height above ground or the boundary-layer depth, or a fitted slope within a tolerance). In the spatial spectra, the modelled slope appears to follow -5/3 towards lower wavenumbers than the left edge of the grey area. Please state the limits, explain how they were determined, and clarify whether the area marks the theoretical inertial subrange or the range where the model follows -5/3. In the temporal spectra, no inertial subrange is marked, although the -5/3 line is shown over the whole frequency range. Please add the corresponding range or explain why it is omitted.

S10. L. 275-281 (Eq. 3) and Fig. B1 (normalisation of the unbiased GPCD)

I am not fully sure that I understand this correctly. As I understand Eq. 3, for a plume that is crossed at the same location by both paths (e.g. close to the release, where the plume is low), it gives twice the vertical column (Ls/sec(SZA) + Lv/sec(VZA) = 2V), whereas Eq. 4 gives V. Close to the source, I would therefore expect the smallest difference between the biased and the unbiased GPCD, since both light paths cross the same plume part. In Fig. B1, however, the difference appears to be largest there. It seems to me that a factor 1/2 is missing in Eq. 3 (or in its implementation). Please check this and clarify whether Eq. 3 is intended to represent the sum of two full-amplitude branches.

S11. L. 424-437, L. 444-446 and L. 533-537 (structure of Sects. 4 and 5, optional)

Large parts of L. 424-437 are literature review (embedding synthetic plumes in observed backgrounds, ML training on synthetic plumes, Bue et al., 2026) rather than a discussion of the results of this study, and they repeat the motivation of the introduction. I suggest moving the example of Bue et al. (2026) to the ML paragraph of the introduction, where it supports the motivation of the paper very well (F1 scores near 1.0 on synthetic data, but below 0.3 on real AVIRIS-NG observations), and shortening this part of Sect. 4 to the implications for the present results. Furthermore, the statement that no reference values or acceptance thresholds exist for multi-metric comparisons at metre scale, and that the presented metrics provide an initial reference point, appears twice (L. 444-446 and L. 533-537). I understand that some repetition helps readers who do not read the whole paper, so I leave it to the authors whether one of the two passages should be shortened.

TECHNICAL CORRECTIONS

T1. General: The font sizes in all figures are too small to be read at print size. Please increase them to a size comparable to the main text.

T2. L. 45: "GHGSat-style" could be replaced by "GHGSat-like" or "GHGSat-type".

T3. L. 79: "MicroHH can represent atmospheric convection [...]" could use "simulate" or "reproduce" instead of "represent".

T4. L. 98-100: The overview of Sect. 2 mentions Sect. 2.5 (slant column densities) before Sect. 2.4 (quantitative metrics). Please swap the two sentences.

T5. L. 115-116: The criterion "plume lengths exceeding 10 m" is used before plume length is defined. Please add a short phrase or a reference to Sect. 2.4.3, where the detection threshold is given.

T6. Fig. 1a: The legend entries "release site (638 px)" and "masked release site (9 px)" are not explained, and the two marker types are hardly distinguishable. Since 638 + 9 = 647 px matches "Plume size: 647 px", the markers seem to show detected plume pixels, while the release site itself is marked by the arrow. Please either explain the difference between the two pixel groups in the caption (e.g. whether the masked pixels are excluded from the analysis) and make it visible in the figure, or use a single marker type.

T7. Fig. 1c: The label "CH4 flux fit" is unclear. It appears to be the mean of the cross-section fluxes, i.e. the estimated emission of 205 +- 20.5 kg/h given in panel (a). Please rename it and state how the value and its uncertainty are calculated (e.g. whether the mean is weighted by the flux uncertainties).

T8. Fig. 6 and L. 237: "3 m-wide" is ambiguous, because "width" is also used for the across-plume standard width of the Gaussian fit. Wording such as "cross-sections spaced 3 m apart" or "3 m-spaced" would be clearer.

T9. L. 258 and Fig. B1: The acronym GPCD stands for "geometry-projected CH4 column densities" in Fig. B1, but for "geometry-projected CH4 column enhancements" in Sect. 2.5. Please use the terms and the acronym consistently.

T10. L. 266-268: The sentence "Close alignment is essential because it exhibited a clear double-plume structure and the modelled and observed plumes were closely aligned" repeats the previous sentence and does not make sense in this form. Please rephrase.

T11. Figs. 5 and A2: The fluxes are plotted in g/s, whereas the release rates in the panel titles are given in kg/h. Please give both units in the titles (e.g. "290 kg/h = 80.6 g/s release") so that the values can be compared directly.

T12. Fig. 8: In panels (b)-(e), the legend lists "CH4" twice (marker and line). Please relabel the entries e.g. as data (retrieved or simulated column) and Gaussian fit. The colourbar of the maps in panel (a) ends at 20 g m^-2, whereas the VCD cross-sections reach about 50 g m^-2. Please extend the colourbar (e.g. with an arrow) to indicate that the colour scale is saturated.

T13. Fig. A2: The legend states "Median", whereas the caption states that the lines show the mean (Fig. 5 uses "Mean"). Please make this consistent.

T14. Supplement, Fig. S2 (and similar figures): In the legend, the markers for MicroHH and AVIRIS-4 are filled, whereas the markers in the panels are open. The red fit line is dashed in the panels, but this line style is not shown in the legend.