Response to Anonymous Referees

We thank Thomas Lauvaux for their helpful and constructive comments and suggestions, which have helped improve the manuscript. Below we provide our answers to specific comments and detail the changes made to the revised manuscript. In the document below, the referee's comments are shown in blue, while the authors' responses are provided in black.

Responses to Thomas Lauvaux

General comment:

The study presents an evaluation of a UAV-based mass balance approach tested over two locations where controlled releases of methane have been performed. The emission rates vary from low (0.02 kg/hr) to large (40.6 kg/hr) amounts, released from single tubes to vertical pipes to ring pipes to long horizontal lines. The sensitivity experiments examine the influence of environmental conditions and flight designs on the final emissions uncertainties. Compared to previous studies, this work brings new insights on the requirements to achieve higher accuracy and precision when using UAV to assess emissions from various types of sources. Overall, the study is fairly complete, with clear explanations regarding the campaign setup and the data analyses, and well-written. The controlled releases have been correctly designed to explore many parameters in real-world applications. Despite a lack of turbulence characterization, which would have been highly valuable to better understand the measured uncertainties, the study remains a significant contribution to this research field. Hence, we recommend this study for publication in ACP once the following points are addressed.

Main points:

1) Causes of the observed biases: While the study has explored some of the potential causes for the observed biases, it remains unclear which ones are responsible for the under-estimation. The absence of turbulence measurements during the flights limits the interpretation of the diagnosed errors in emissions. It is understandable that the additional cost of a 3D sonic is far from negligible, but considering your conclusions (wind direction variability, distance to the source), it seems obvious that a turbulence analysis would have helped you diagnose the relation between advection and turbulence, and their potential impact on the retrieved emissions.

While the authors have noted the relation to the wind direction variability, one can wonder if the variability arises from the turbulence at high frequency or to slower motions of the plume due to large-scale eddies. This parameter is critical to determine the most favorable environmental conditions for future campaigns, as the motion of the plumes within the perpendicular plane is proportional to the mean wind speed and to the turbulent structures. An additional 3D sonic anemometer would have provided additional insights on the size and impact of eddies, especially when flying near the source. We recommend that you discuss that point, and to the extent possible, you could use the 2D sonic data to examine the frequencies of wind direction changes over the sampling periods.

We agree that vertical winds would have been useful to gain insights into not only turbulence, but also plume lift rate with distance downwind, which can become important at longer ranges downwind in convective conditions. However, the mass balance method is concerned with flux through a 2D horizontal plane, and thus it is the average horizontal wind vector that is relevant to the method's formalization. Variability in horizontal wind can help with error characterisation as we discuss in the paper and prior work cited. But vertical winds would be helpful more generally for sure. However, It is not possible to measure vertical winds reliably on a rotary UAV platform due to the physical obstruction of the UAV below (if mounted above the rotors). Turbulence is complex to measure and model, particularly at the fine spatial and temporal scales relevant to plume meandering within a UAV measurement plane. As the referee mentions, robust characterisation typically requires high-frequency 3D wind measurements, which are difficult to obtain from a moving airborne platform and are further complicated by rotor-induced flow disturbance. Given this complexity, we were not able to directly measure or model turbulence at the scale needed to fully diagnose its relationship with the retrieved emissions in this study.

We note, however, that large-scale eddies are not expected to introduce a systematic bias in the mass balance flux estimate, since they primarily rotate mass within the sampled plane rather than transporting it net across the plane boundary: an eddy-driven increase in along-wind velocity or concentration in the upper part of the measurement plane tends to be accompanied by an opposing motion in the lower part, such that their net contribution to the integrated flux is expected to largely cancel when averaged over the full plane. This is the assumption always made in horizontally-advictive mass balancing and used by many teams internationally. But it's important to acknowledge its potential effect and account for it by uncertainty budgeting, which is what we do when considering horizontal wind variability (which captures the effects of eddies crossing the advective sampling plane. We would therefore expect large-scale eddy structures to contribute primarily to random scatter across individual flights, consistent with the improvement in RMSE and correlation we observe after averaging (Fig. 4c, d), rather than to the systematic bias itself.

Averaging to reduce the bias: Based on Figure 2, and as observed by most UAV studies, the motion of the plume due to passing eddies results in random intercepts of the CH4 plume by the drone, at various locations across the flight plane. Clearly, the displacement of the observed peaks is a testimony of the plume fluctuations driven by turbulent structures. Hence, as one collects data over multiple flights, it may reduce the bias (CEDAR) or not (Bedford). The distance to the source becomes critical as one observes more fluctuations near the release location, thanks to diffusion. Assuming that the method is unbiased, averaging is then recommended to reduce the measured errors. Therefore, it would be more valuable to discuss how long (or how many flights/transects) are needed to reach the minimum level of uncertainties (bias).

The reviewer's understanding is correct. Averaging helps to cancel out these effects - so long as distance to source remains broadly the same. This is a key assumption of mass balancing that we follow when designing flights. Literature also provides useful information relative to the number of flights (sampling time) needed in different operational contexts. For example, Fosco et al. (2024) and Yong et al. (2024), cited in our manuscript, noted that a single UAV survey may not capture the average plume morphology as they are snapshots in time. Therefore, repeated flights are recommended to reduce spatial and temporal variability. Scheutz et al. (2025), using a similar UAV-based mass balance method, provide quantitative evidence for this recommendation. In their study, they found that increasing the number of flights reduced the maximum quantification error from approximately 14% to approximately -3%, when at least three flights were averaged, with the standard error of the mean decreasing to 0.3 kg h^{-1}.

This is consistent with results already presented in our manuscript (Sect. 3): when flights are grouped by release rate and averaged (Fig. 4c, d), agreement improves substantially at both sites. At Bedford, correlation increases to $r = 0.994$ with a similar slope (0.68) and RMSE decreases to 38.4%, though bias remains comparable (-36.5%); at CEDAR, the averaged results show near-perfect agreement (slope = 1.01, $r = 0.94$), with very low bias (-1.63%) and RMSE (19.9%), and all controlled release rates are captured within the 1σ uncertainty ranges of the averaged estimates. As already noted in the manuscript, this improvement is consistent with previous studies indicating that a minimum number of repeated flights per release rate (e.g. $n \geq 3$, Scheutz et al., 2025) enhances representativeness, with the reduced scatter and increased correlation after averaging suggesting that individual flights are influenced by environmental variability, while aggregating multiple flights provides a more robust flux estimate.

We would also like to distinguish two related but separate issues raised by the reviewer's comment: the number of repeated surveys needed to average out random, turbulence-driven variability, and the vertical/horizontal coverage needed within a single survey to avoid systematically missing part of the plume. The latter is a known source of bias that averaging alone cannot correct in principle, and is already discussed in our manuscript with reference to Morales et al. (2022), who found that failure to capture the true release in a subset of their flights was linked to incomplete vertical/horizontal plume capture, as well as via our own near-ground extrapolation correction (Sect. 2.2) and the recommendation to sample further downwind or as low as possible (Sect. 3). In this study, adequacy of vertical/horizontal coverage was assessed post-flight by reviewing the recorded concentration data. Real-time in-flight monitoring of plume boundaries would be a valuable addition for future campaigns. We note that our controlled-release flights (single battery set, given the simpler, single ground-level source) required less vertical extent than our operational landfill surveys (two flights per survey, up to ~120 m), reflecting differences in source complexity between these applications.

To address this comment, we added the following discussion in the revised manuscript to better discuss this important comment and information available in our study and in the literature (line 282-287, page 12):

This is further supported by Scheutz et al. (2025), who found that increasing the number of flights reduced the maximum quantification error from approximately 14% to approximately -3% once at least three flights were averaged, with the standard error of the mean decreasing to 0.3 kg h^{-1} by eight flights. It is worth noting that repeated averaging reduces random, flight-to-flight variability but does not correct systematic bias arising from incomplete vertical or horizontal plume capture within a single survey (Morales et al., 2022), which we address separately via the near-ground extrapolation correction (Sect. 2.2).

Low-level transect: The authors have highlighted the importance of the lowest transects when calculating the overall mass flux. However, the method of interpolation (and more specifically extrapolation near the ground) used here was never discussed nor tested. While these methods are widely used to interpolate 2D fields, the mass conservation is not guaranteed and can pose significant challenges in such applications. A brief examination of the interpolation method would be beneficial to better understand if the interpolation/extrapolation technique is not causing additional positive biases, compensating for the lack of data (plume intercepts).

We agree mass conservation is not guaranteed in any extrapolation, which is why we only add this to the uncertainty budget, and not to the flux estimate (see also comment to Review 1 on this). We believe this is a very important distinction and we agree with the reviewer. To provide more information on the method, the near-ground extrapolated flux is calculated by assuming a constant flux density, equal to the mean flux density measured across the lowest sampled transect, applied uniformly over the unsampled region between the lowest sampled height and the ground. We chose this simple, constant-value assumption because our method is applied across a range of source types and site conditions, and we wanted to avoid introducing additional profile parameters (e.g. a decay rate or roughness length) that would need separate justification or calibration in each case, without concurrent near-surface measurements available to constrain them. Given this lack of ground-truth data, we consider a low-parameter, transparent assumption preferable to a more complex profile whose accuracy we would be unable to verify.

For this reason, we do not add the extrapolated contribution to the central flux estimate, but instead incorporate it only into the positive uncertainty bound, since it represents an expected under-bias due to incomplete vertical sampling (Sect. 2.2). This design choice means our approach cannot overestimate emissions by construction, since the central estimate remains based solely on directly measured concentrations. We find support for this approach in our own results: including the extrapolated term in the upper uncertainty bound increases the fraction of cases in which the known release rate falls within the calculated uncertainty range from 46% to 72% (Sect. 3), indicating that the resulting uncertainty range better represents the true variability in emissions than when the extrapolated contribution is excluded.

To action the reviewer's comment, we have revised the manuscript to state this extrapolation method in more detail - thank you for the recommendation. Future work could test and refine this assumption directly by incorporating near-surface measurements alongside UAV surveys, for example using ground-based sensors or robotic platforms to sample the near-ground layer that the UAV cannot access.

We added the following lines in the revised manuscript (lines 163-168, page 6):

The near-ground extrapolated flux was calculated assuming a constant flux density, equal to the mean flux density of the lowest sampled transect, applied uniformly to the unsampled region below it. This simple assumption avoids introducing additional profile parameters (e.g. a decay rate or roughness length) that would require separate justification across the range of source types and conditions to which this method is applied, and which cannot be constrained without concurrent near-surface measurements. Future work could test this assumption directly, for example using ground-based sensors or robotic platforms to sample the near-ground layer.

Technical comments:

Line 325: How did you determine the atmospheric conditions?

Atmospheric stability was assessed after the campaigns using ERA5 reanalysis (cloud cover, incoming shortwave radiation and boundary-layer height where available), averaged over the 08:00-16:00 measurement window, at each campaign's coordinates and dates. Bedford (15-16 May 2024) showed consistently high cloud cover (89-99%) and low radiation (247-361 W m^{-2}), consistent with neutral, cloud-dominated conditions. Boundary-layer height information was not available for this site. At CEDAR (8-10 September 2025), conditions shifted during the campaign: 8 September was largely clear and highly radiative (21% cloud cover, 437 W m^{-2}) with a deep boundary layer (1148 m), consistent with convective conditions. By 10th September, cloud cover increased to 82%, radiation dropped to 234 W m^{-2} and the boundary layer had swallowed to 807 m,pointing to a swift towards neutral conditions.

We have added the clarification to the manuscript (Line 395-398, page 19) and note it reflects gridded reanalysis averaged during the field-measurement window :

During the controlled release experiments, atmospheric conditions at both sites varied between neutral-buoyant and convective-buoyant, as inferred from ERA5 reanalysis cloud cover, incoming shortwave radiation and boundary-layer height (where available) for the corresponding dates, coordinates and measurement hours (8:00-16:00) (Hersbach et al., 2023; accessed via the Open-Meteo historical weather API, Zippenfenig, 2023).

Line 331-336: the ratio between turbulence and advection is likely to define the variability, which can explain why one of the two alone is insufficient to explain the discrepancies.

We agree that the ratio between turbulence and advection could play a role in this variability. This is because turbulence intensity tends to decrease as wind speed increases, which is consistent with our finding that bias decreases with increasing wind speed: stronger winds leave less time for the plume to spread laterally, so directional variability likely matters relative to the wind speed rather than on its own. However, our anemometer only measures wind in 2D, so we could not directly assess turbulence or quantify the suggested ratio in this study.

We added the following to the manuscript to explain this (line 407-411, page 19):

This is also consistent with the earlier finding that method performance improves (bias reduces) with increasing wind speed: stronger wind speed leaves less time for the plume to spread laterally, so the variability of the wind direction likely matters relative to the wind speed rather than on its own. We note that our anemometer only measures wind in 2D, so we could not directly assess turbulence. Future work could include a 3D anemometer to check if the ratio between turbulence and advection affects this variability.

Figure 2: Looking at the highest transect, it seems quite obvious that you have missed the vertical extent of the plume motions. While you discuss the impact of missing low-altitude transects, you never discussed the potential lack of high-altitude transects.

Indeed our discussion is mostly focused on the incomplete vertical sampling for the near-surface area. This is because methane sources are located on the surface and therefore the concentration in the area above the surface is expected to greatly affect the flux quantification. In contrast, in higher altitudes methane concentration is expected to be lower than above the surface, so missed fluxes near the top of the measuring plane are expected to be smaller in magnitude than near-surface losses. Nevertheless, we agree this contribution is not necessarily negligible, particularly under strong convective conditions or for elevated sources in real-world conditions. However, this would be impossible to attempt to account for in error budgeting (unlike the near-surface layer, which has a known finite distance). We have added the following discussion to the manuscript (line 179-186, page 7):

A similar, though less frequent, limitation might apply to the top of the measurement plane. Since methane sources are located at or near the ground level, concentrations are expected to decay with altitude so missed flux above the highest transect is generally expected to be smaller than near-surface losses. We reviewed all flights and found this occurred only in a small minority of cases, likely due to a low flight ceiling (e.g. Fig. 2 where the highest transect reached 12 m). No correction was applied for this effect in the flux estimates or its uncertainties. We recommend monitoring real-time CH_4 *concentration during flight to confirm the highest transect has returned to background levels, and using a taller vertical sampling range where operationally feasible, particularly under convective conditions or for elevated sources, when vertical plume lofting is more likely.*

Figure 5: the separation by emission rates could also be seen as a split between the two sites, or by distance. You should indicate these important facts to clearly separate if the errors scale by source level or due to other parameters.

We agree that the release rate bins in Figure 5 correspond closely to site (low rates = Bedford, high rates = CEDAR) and, to a lesser extent, to distance from source (slightly shorter distances at Bedford, slightly longer at CEDAR). Site-specific results are included in Fig. 4 and Table 2 in the submitted manuscript. Figure 5 instead presents a high-level, cross-site summary, organised by release rate, rather than site, because we consider rate-dependence discussion a more generalizable message that is relevant to the wider scientific community.

Regarding distance, our results show a negligible correlation between downwind distance and absolute bias for the full dataset. This means that it is unlikely the observed pattern in Figure 5 may be driven by the distance from the source alone. Regarding the sites, we chose not to split the analysis by site, as this would leave too few points per group for statistically robust results (a concern the reviewer rightly raises for Figure 7 below). We instead discuss bias in terms of release rate, since there is a clear underlying reason for the trend: a given absolute error represents a much larger relative (%) bias for a small release than for a large one. This provides a site-independent explanation for the observed pattern, which we consider more informative than a site- or distance-based grouping. We have revised the manuscript accordingly (lines 309-314, page 14):

Release rate groups in Fig. 5 correspond closely to site and less strongly to distance from source. Distance from source shows a negligible correlation with absolute bias across the dataset (see Sect. 3.2), making it an unlikely driver for the observed pattern in Fig. 5. We do not split the analysis by site, as this would reduce statistical robustness given the relatively limited number of flights per group. We instead interpret the trend in terms of release rate, since a given absolute error represents a larger relative (%) bias for smaller releases, which is a site-independent explanation for the observed patterns.

Figure 7: Some of these regressions are far from convincing, with some of the lines being decided by few points (e.g. distance). The overall fits are statistically limited by the lack of points. Add some statistical metrics to evaluate the goodness to the fit, and add a discussion about the level of confidence for each variable.

To address this comment, we computed the p-value and 95% confidence interval of the regression slope for each variable in Figure 7, and added shaded 95% confidence bands to the figure. These statistics (R^2 and p-value) are now reported for each variable throughout Sect. 3, and a summary discussion of confidence levels across all variables has been added (lines 432-437, page 20) as follows:

Overall, average wind speed, $\Delta\theta$ and the percentage of the measurements within $\Delta\theta$ show statistical significant relationships with bias ($p<0.05$), while release rate shows a trend that does not reach statistical significance ($p=0.10$), and wind direction standard deviation and distance show no significant relation ($p=0.76$ and $p=0.79$, respectively). The corresponding 95% confidence intervals, shown in Fig. 7, are widest for variables with sparser sampling at the upper end of their range, indicating that these fits should be interpreted with caution.

References

Fosco, D., De Molfetta, M., Renzulli, P., and Notarnicola, B.: Progress in monitoring methane emissions from landfills using drones: an overview of the last ten years, Science of the Total Environment, 945, 173 981, https://doi.org/10.1016/j.scitotenv.2024.173981, 2024.

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., and Thépaut, J.-N.: ERA5 hourly data on single levels from 1940 to present, ECMWF, https://doi.org/10.24381/cds.adbb2d47, 2023.

Morales, R., Ravelid, J., Vinkovic, K., Korben, P., Tuzson, B., Emmenegger, L., Chen, H., Schmidt, M., Humbel, S., and Brunner, D.: ´ Controlled-release experiment to investigate uncertainties in UAV-based emission quantification for methane point sources, Atmospheric Measurement Techniques, 15, 2177–2198, https://doi.org/10.5194/amt-15-2177-2022, 2022

Scheutz, C., Knudsen, J. E., Vechi, N. T., and Knudsen, J.: Validation and demonstration of a drone-based method for quantifying fugitive methane emissions, Journal of Environmental Management, 373, 123 467, https://doi.org/10.1016/j.jenvman.2024.123467, 2025.

Yong, H., Allen, G., Mcquilkin, J., Ricketts, H., and Shaw, J. T.: Lessons learned from a UAV survey and methane emissions calculation at a UK landfill, Waste Management, 180, 47–54, https://doi.org/10.1016/j.wasman.2024.03.025, 2024.

Zippenfenig, P.: Open-Meteo.com Weather API, Zenodo, https://doi.org/10.5281/ZENODO.7970649, 2023.