We wish to thank both reviewers for their insightful and helpful comments, their attention to the details of the analysis, and their positive feedback about the manuscript. We have incorporated all the suggestions where possible, answered all the reviewers' questions, and added new analysis based on their suggestions. We believe this has improved the manuscript considerably.

In the attached file, we respond to each of the reviewer comments. In cases where we refer to added or modified text from the revised manuscript, the text from the manuscript is in quotes and the new text is italicized.

RC1: 'Comment on egusphere-2025-4542', Joseph Pitt, 01 Dec 2025 reply

This study investigates sources of error in typical aircraft mass balance experiments and provides guidance for experiment design. It is well written and easy to follow, with clear conclusions: storage leads to random error in E_H/E_S, whereas extrapolation below the lowest transect can be an important source of systematic error. This is consistent with previous studies, but the thorough investigation presented here provides valuable insight for planning and evaluating future real-world data. I suggest that this paper is suitable for publication with only minor revisions. I think it would benefit from a slightly expanded discussion on the points below.

There is currently very little mention of the background concentrations. I understand that this is not a factor in the simulated data, as all the tracers released in the model come from sources within the domain. However, in real world examples variability in the background can be an important source of error, so at least some discussion of this is required. In particular, it impacts statements such as that in L368-370. Going further downwind may reduce the sources of random error addressed here, but there is a trade-off in terms of signal-to-noise above background.

Following the statement (end of 5th paragraph in Section 3.1.1, previously L370) we add the following text "*In a real-world scenario, there may be additional error due to noise in the concentration measurements, particularly for concentrations with a high background level. In these cases, moving further downwind (where the concentration enhancements above background due to the plume are much smaller), may result in a relative increase in error.*"

Further, at the end of the 3rd paragraph in the Conclusions, we add "*Generally, in real-life conditions, any reduced uncertainty due to flying further downwind from source must be balanced against increased relative uncertainty due to instrument measurement noise, especially for pollutants with high background concentrations (due to the weaker concentration signal as the plume disperses).*".

It is interesting that the kriging interpolation resulted in an overestimation of the instantaneous screen (L375). It would be great to see some more investigation of this. Was anisotropy in the variogram considered? I wonder if the variogram becomes more isotropic as you move further from the source? That would make intuitive sense to me. Were other functions (i.e. other than the spherical function mentioned) tested when fitting the variogram? It could also be interesting to see if this choice impacts the overestimation, although I appreciate that it is hard to draw general conclusions because the best function will always be specific to an individual flight. The same goes

for the area source flights – L487 points to the kriging as a potentially significant error source so it would be good to see this case investigated too.

We have added new subsections 3.1.3 (for stack sources), and 3.2.3 (for area sources). The new sections are titled "Optimizing Screen Interpolation". Our test (not included) demonstrate that some flights do show anisotropy in the variogram, but the end results are not sensitive to the range of the variogram fits. For example, halving or doubling the model range value changes the average concentration in the screen by less than 2%. Hence, accounting for anisotropy would have little effect here.

As part of this new analysis, we compared results with spherical and exponential kriging, as well as Voronoi nearest-neighbour interpolation. Results are not sensitive to the interpolation model type.

The new subsections are as follows:

"3.2.3 Optimizing Screen Interpolation

We repeat the investigation of the interpolation described in Section 2.4.3 and 3.1.3 for the small and large area sources. Here we only investigate the kriging interpolation with the spherical variogram model, having demonstrated small differences between the different interpolation methods in Section 3.1.3. Results for the small area source are similar to the results for the stack sources shown in Figure 11. Close to the source ($D = 2$ km), there is overestimation in the interpolated concentration for $z > 150$ m, and significant underestimation of the extrapolated concentration for $z < 150$ m. Further from the source ($D \geq 6$ km), the average ratio (for each flight set) of interpolated to actual concentration ($\langle C_K \rangle / \langle C_M \rangle$) for $z > 150$ m varies from 0.96 to 1.04, while the extrapolated concentration below 150 m shows significant errors (with $\langle C_K \rangle / \langle C_M \rangle$ as high as 1.22). In all cases, the variability between flights decreases with downwind distance, ranging from 4 to 7% for the $z > 150$ m interpolation for $D \geq 6$ km. For the large area sources, there is much less error in the interpolation above 150 m, relative to the error in the interpolation from the stack or small area sources. Average values of $\langle C_K \rangle / \langle C_M \rangle$ range from 0.99 to 1.02, and the variability between flights is less than 4% for all downwind distances. There is significant underestimation of the extrapolated concentration below 150 m, especially for the Sep 2 flight sets, which was previously demonstrated in Section 3.2.1 (e.g. Fig. 4k)."

"3.1.3 Optimizing Screen Interpolation

As shown in Figure 3, the estimation of E_S is generally higher in the non-instantaneous flight sets, relative to the instantaneous flight sets, for $D \geq 6$ km, but lower in the non-instantaneous flight sets for $D = 2$ km. This may be due to the kriging interpolation causing an overestimation of the screen concentrations. To test the interpolation, we sampled the instantaneous screens using the non-instantaneous flight path positions, allowing us to compare the interpolated screens with the high-resolution model output screens. We applied the same extrapolation of a constant concentration for heights below 150 m. The average concentration is then calculated from the interpolated screens, $\langle C_K \rangle$, which can be compared with the average concentration for the model output screens, $\langle C_M \rangle$. The ratio of $\langle C_K \rangle$ to $\langle C_M \rangle$ is shown in Figure 6, separated into the averages above 150 m and below 150 m (where the concentration is assumed constant). For simplicity, we only compare 2 cases: Aug 20 and Sep 2 with 16:20 flight start times.

The differences between the different interpolation methods are small relative to the errors in the interpolation at different downwind distance, D. For $z > 150$ m, kriging with the spherical variogram model gives a slightly better average ($<C_K>/<C_M> = 1.07$ for all downwind distances on Aug 20 and $<C_K>/<C_M> = 1.01$ for Sep 2, versus 1.08 and 1.01 respectively for an exponential variogram model, and 1.10 and 1.03 respectively for the Voronoi nearest neighbour). For the spherical model, the results are also not sensitive to the goodness of fit of the variogram model. For examples, for the Aug 20 values for $z > 150$ m, halving or doubling the range value of the variogram model changes the average value of $<C_K>/<C_M>$ by less than 2%. Hence, the choice of interpolation method and the details associated with those choices seems to be less consequential than the changes in the sparseness of the sampling at different downwind distances.

The interpolation for $z > 150$ m generally overestimates the actual concentration and shows high variability between flights when close to the source ($D = 2$ km). Further downwind ($D \geq 6$ km), the interpolation is significantly improved and the variability between flights is reduced, with values of $<C_K>/<C_M> = 1.00$ and $\sigma = 6\%$ for the Aug 20 16:20 flights, and $<C_K>/<C_M> = 1.01$ and $\sigma = 3\%$ for the Sep 2 16:20 flights (both at $D = 10$ km with the spherical kriging).

The total interpolation errors appear correlated with the difference between the instantaneous and non-instantaneous flight set emissions estimates shown in Figures 3a and c. For example, in Figure 3a, for the Aug 20 16:20 flight set at $D = 2$ km, $E_H/E_S = 0.63$ for interpolated, non-instantaneous flight set, compared to $E_H/E_S = 0.96$ for the instantaneous flight set. Figure 6 for the same flight set (at $D = 2$ km) shows an underestimation with $<C_K>/<C_M> = 0.43$ for $z < 150$ m and $<C_K>/<C_M> = 1.19$ for $z > 150$ m, suggesting a net underestimation (although the relationship between concentration and advection flux is also influenced by wind speed and the plume maybe not be evenly distributed between below and above 150 m). Similarly, for most other distances shown in Figure 6, an underestimation (or overestimation) in $<C_K>/<C_M>$ is generally associated with a similar scale underestimation (or overestimation) in E_H/E_S for the non-instantaneous flight sets relative to the instantaneous flight sets (which are not interpolated). This implies that the sparseness of sampling is a significant source of error for interpolation close to the stack sources ($D = 2$ km), while further downwind ($D \geq 6$ km), there can be significant errors due to the extrapolation of a constant concentration below 150 m (as discussed in Section 3.1.1 and shown in Figure 4).”

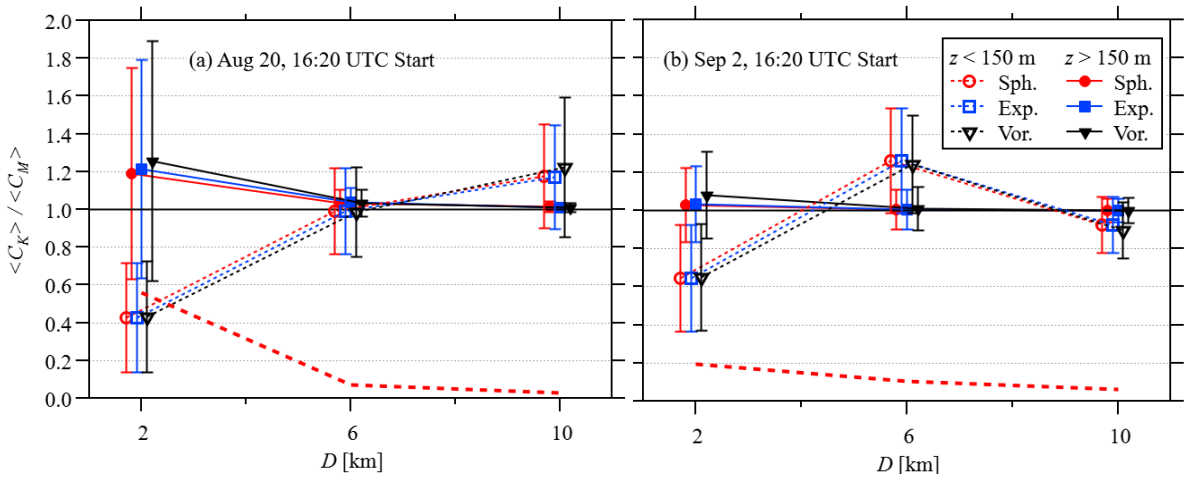


Figure 6. A comparison of average concentration from the instantaneous screens (at downwind distance, D) $<C_M>$, to the average concentration from an interpolation of those flights with sparse sampling $<C_K>$ downwind of stack sources (given as ratio $<C_K> / <C_M>$). The averages are separated into below and above 150 m (open symbols with dotted lines, and closed symbols with solid lines, respectively), where the below 150 m concentrations are assumed constant (see Fig 4). Three interpolation methods are compared: kriging with a spherical variogram model (red circles), kriging with an exponential variogram model (blue squares), and Voronoi nearest-neighbour (black triangles). The markers are offset slightly for clarity. The standard deviations of the 10 flights (σ) are shown as error bars, as well as red dashed lines for the spherical kriging (> 150 m) only. Results are shown for the Aug 20 (a) and Sep 2 (b) 16:20 flight sets, corresponding to Fig. 3a and c, respectively.

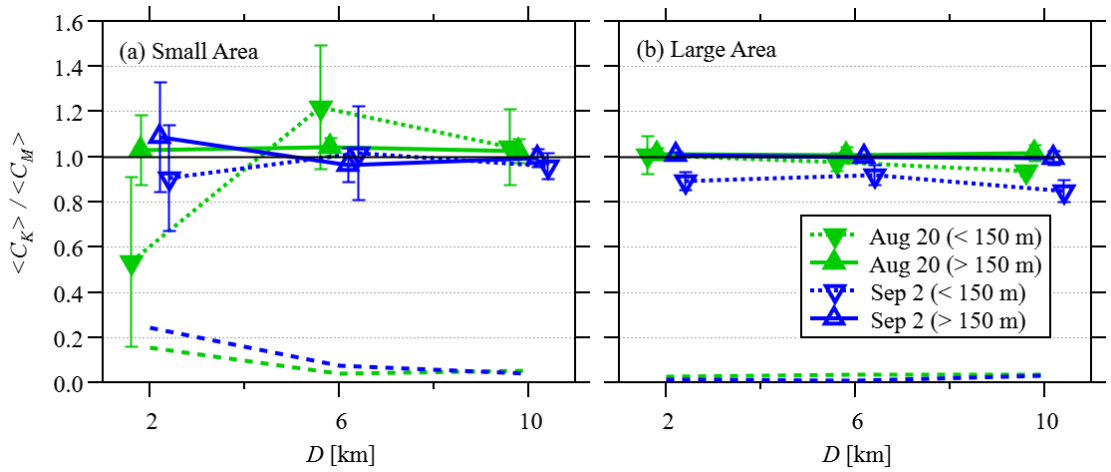


Figure 11. A comparison of average concentration from the instantaneous screens (at downwind distance, D) $<C_M>$, to the average concentration from an interpolation of those flights with sparse sampling $<C_K>$ downwind of stack sources. The averages are separated into below (downward triangles with dotted line) and above (upward triangles with solid lines) 150 m, where the below 150 m concentrations are assumed constant (see Fig 4). The markers are offset slightly for clarity. Results are shown for (a) the small area sources and (b) large area source, for the Aug 20 (green solid triangles) and Sep 2 (blue open triangles) 16:20 flight sets. The standard deviations of the 10 flights (σ) are shown as error bars, as well as dashed lines (for > 150 m only)."

The investigation of the vertical transect spacing is interesting but the results are hard to interpret. The hypothesis that plume movement could be responsible for the changes seen in the Sep 2 case seems plausible, but it would be nice to see this tested. Seeing as we are dealing with simulated flights, could a test be done where the order of the transects is changed?

We have done this analysis for both flight sets (Aug 20 and Sep 2). The figure now includes both cases in reverse, and the following text is added to the end of Section 3.1.4, "*To investigate this, we repeat the flight sets using the same flight paths with the directions reversed (i.e. top-to-bottom). In these cases, the flight begins at 16:20 (for the first flight in the sets) at the highest point and each flight samples at identical locations in the opposite direction, finishing at the lowest point at a height of 150 m. The reversed direction results in significant improvement in the estimation of E_H/E_S for the Aug 20 flight sets (especially at smaller transect spacing), but increased error for the Sep 2 flight sets. In both cases, the variability between flights in each set is reduced for a transect spacing*

of $\Delta Z = 50$ m, but is slightly higher than the variability of the bottom-to-top flight sets with $\Delta Z \geq 100$ m. As mentioned in Section 2.4, real flights are often flown in an upward direction so that the vertical extent of the plume can be determined while flying. While it is difficult to know the vertical extent of the plume beforehand (which would be required for a flight in the downward direction), these results demonstrate the potential advantage of flying once in the upward direction, followed by a subequent flight back in the downward direction."

L213/216 – refers to the known emissions as Es but I don't think this has been defined yet

In both cases we add "the *ratio of the storage term to the known emission rate* (S/E_S)."

L225 – in some cases even faster than 2 Hz. I know the UK FAAM aircraft has a CO2/CH4 LGR with a data acquisition rate of 10 Hz, although the cell turnover time means that the effective frequency of measurement is less than this (more like 7 Hz I believe).

This is modified to "…from 0.5 to *as high as 10* Hz *(e.g. France et al., 2021)*," and the citation for the France et al (2021) study is added.

L360 – it might be worth rephrasing this to clarify that it is E_H/E_S which is lower in the non-instantaneous cases (i.e. the underestimation is worse). A "lower underestimation" could perhaps be misinterpreted.

We have modified the text to "(*i.e. a lower E_H/E_S value is estimated by the assumed below 150-m concentration relative to what would be determined with the actual below 150-m concentrations*)".

L413 – typo "sometimes"

Corrected.

Figure 9 – formatting error on some axes labels

This error occurred in the Word to PDF conversion. We have reformatted the figure so that this doesn't happen.

L576-577 – it makes qualitative sense that more information below the lowest transect would help. Could this be tested? At least for the case of a mobile vehicle you could presumably add an extra transect at z=0 with a typical vehicle speed and see what difference this makes

We have added new sections to investigate this possibility, below is newly added Section 2.7 and subsection 3.1.5, and the modification to Figure 3.

"***2.7 Ground-vehicle Sampling***

We also investigate the potential improvement to the emissions estimate through ground-based mobile vehicle concentration sampling. For the case of the Aug 20 (starting at 16:20) flight set downwind of the stack sources, we simultaneously sample concentrations at the lowest model level beneath the flight path. Although vehicle path locations are typically limited to roadways, we investigate here the highly idealized case where a car or truck can drive directly beneath the flight path of the aircraft for the duration of the flight. We assume a constant vehicle speed of 60 km/hr (16.7 m/s) and drive a single transect south-eastward from the most NW location. These values are

then used in the interpolation of the screen (at $z = 1$ m) without the need to assume a profile below a height of 150 m. Results are discussed in Section 3.1.5."

"3.1.5 Adding Ground-based Vehicle Measurements

Figure 4 demonstrates that there can be considerable error associated with the assumed concentration profile below the lowest flight path of 150 m. As discussed in Section 2.4, in some situations, it may be possible to measure concentrations at ground level with a mobile measurement platform on a car or truck. Results for the Aug 20, 16:20 start flight set augmented by ground-based vehicle measurements are shown in Figure 3a. With these surface-level measurements, the assumption of a constant profile below the height of 150 m is no longer necessary, and the screen can be interpolated using these additional measurement values. The horizontal advective flux with ground-based measurements consistently overestimates the emission rate for all downwind distances, with values of E_H/E_S ranging from 1.14 to 1.30 for $2 \leq D \leq 12$ km. This demonstrates that the underestimation of the horizontal advective flux close to the stacks ($D < 6$ km) with an assumed constant concentration below 150 m is predominantly due to a large amount of the plume being below the lowest flight path, as was shown in Figure 4. Further, the variability between flights is reduced by up to 6% (at $D = 4$ km). The results demonstrate significant value is obtaining surface-level measurements where possible, so that extrapolation below the lowest flight path is not required."

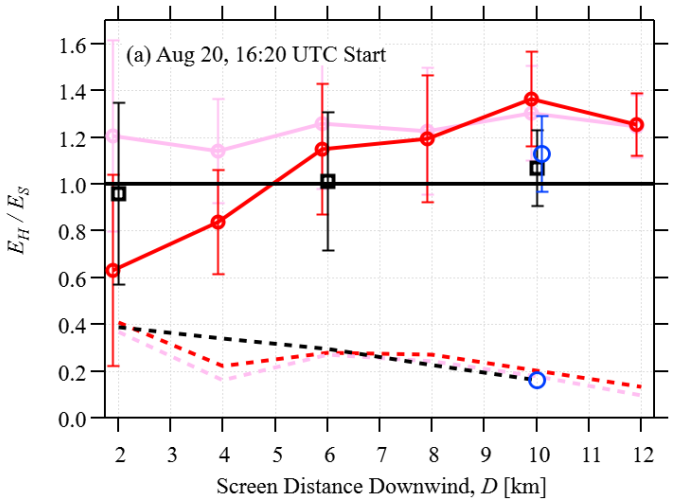


Figure 3a Modification: The following text is added to the figure caption "*The pink lines in (a) demonstrate a case with additional measurements from a ground-based vehicle, discussed in Section 3.1.5.*"

RC2: 'Comment on egusphere-2025-4542', Anonymous Referee #2, 18 Dec 2025 reply

Review of: “Optimizing Airborne Emission Rate Retrievals with Sub-Hectometre Resolution Numerical Modelling”, by S. Fathi, M. Gordon & J. Hao.

By: Anonymous Reviewer

General comment:

The manuscript presents a detailed model-based study aimed at providing insights into recommended strategies for flight planning when employing mass-balance methods. It builds upon earlier works (both modelling and measurements) and is focused on various source types (dispersed area sources and tall stacks) around the Canadian Athabasca oil sands, where heavy oil industry is responsible for releases of large amounts of various atmospheric pollutants. The primary tool used in the investigation is the WRF model run at very high spatial resolution (50 m horizontal). Based on the setup evaluated in previous studies, the model delivers highly resolved spatial concentration fields based on the assumed source distributions, which form the basis of the analysis. The authors evaluate the ability to retrieve true source emissions from the measurements performed using hypothetical airborne platforms (either aircraft or UAVs), represented by extracting model-predicted fields at locations and times mimicking a way airborne platforms would be normally flown. The authors evaluate accuracy of the estimated emissions varying factors like flight strategy, measurement distance from the source and the role of vertical data density on the said accuracy, including data density. They interpret this data in order to search recommended ways to sample similar emission sources in real-world conditions.

I find the paper extremely interesting, well written, and generally well structured, with minor editing notes listed at the end of this review. I find the quality of the modelling work very high and the interpretation of data shows very good understanding and the topic. However, I also identify several major flaws that need to be addressed before the paper can be considered for publication.

PS – I ask authors not to be discouraged by multiple remarks. These are meant to be constructive and I want to underline that I think the work is of high quality and will make a valuable scientific contribution when the following concerns are addressed.

Specific major comments:

1. I believe the use of statistics needs to be reworked. The authors incorrectly describe uncertainties, using S.E. = 1.96 $\sigma / \sqrt{n}$, when S.E. should be defined without the 1.96 factor. The intent is right, but according to metrology standards this quantity is correctly “expanded standard uncertainty of the mean”, or “expanded standard error” (although this is discouraged). See JCGM 2008 for details. Here also k (coverage factor) is chosen inappropriately – 1.96 is the correct value when the effective degrees of freedom are extremely large. In their study the authors have only 10 repetitions, in which k value will be higher than 2 (if the results were uncorrelated – see below).

All references to the standard error have been removed. The value of k is now used which reflects the degrees of freedom, determined from the effective number of observations, discussed in point 2a below.

2. Following to the above, but much more important, is that the authors incorrectly assume that the observations within their subsets are independent, and ignore the existence of correlation. In

fact it was demonstrated in the past (Gerbig et al., 2003) and in more recent works (Fuentes-Andrade et al. 2024, Galkowski et al. 2025) that the atmospheric signals of atmospheric pollutants are correlated at spatial and temporal scales large enough to be of significance to the measurements like those investigated in this work. In Galkowski et al., CO_2 emissions from elevated stacks were found to be auto-correlated down to a distance of 4 kilometers with persistent spatial (mostly horizontal) structures affecting plumes down to distances of even 20 kilometers. Although here the correlated structures are likely to be shorter (smaller PBLH, lower emission altitudes), the authors cannot ignore correlation in the signals in their analysis. Formally evaluating and including impact of correlation are likely affect the results in two major ways:

a) the uncertainty ranges calculated for the emission rates are expected to increase, as the effective degrees of freedom (number of independent measurements) will be reduced for each distance, for which emissions were evaluated. For methods on evaluating degrees of freedom, see e.g. works cited above.

We calculated the effective number for each flight set following Zieba (2010), which gives an average n_{eff} of 8.0 for the stack sources, 7.9 for the small area sources, and 7.5 for the large area sources. Given the importance of this analysis, we have added a subsection to discuss the calculation of uncertainty as follows:

"2.4.2 Uncertainty Estimation

Through the statistical analysis of multiple flights, we can also assess how effective repeated flights (or multiple sampling with 2 or 3 UAVs or aircraft) are in reducing the measurement uncertainty in the emission rate estimate. Since subsequent flights may not be statistically independent, we determine the autocorrelation function of the times series of the horizontal advective flux (E_H) for each set of 10 flights. This is used to calculate the effective number of flights, n_{eff}, following Zieba (2010), which is < 10 if subsequent flights are not statistically independent. This gives the effective degrees of freedom for the calculation of the mean (n_{eff} – 1), which can be used to calculate the expanded uncertainty of the mean, following JCGM, 2008. If we can assume that the variability within a flight set (σ) is representative of the real variability a flight would encounter under similar conditions, then we can use that value to estimate the uncertainty in a single flight estimate of E_H in the real world. Using the value of $n_{eff} = 7$ as an example, we are 95% certain that a single estimate is within 2.45σ of the actual mean E_H value (Table G2 in JCGM for 6 degrees of freedom and a 95% confidence interval). If, for example, two real flights can be flown (far enough apart in time to assume they are independent measurements), then this uncertainty in the estimate of E_H is reduced by $\sqrt{2}$ to 1.41σ. It is noted that the uncertainty calculated here effectively combines our uncertainty in the variability in the flights due to a limited number of samples (approximately 26%, since an infinite number of flights would reduce 2.45σ to 1.95σ) with the actual variability between flights, which could be due to storage fluctuations, interpolation/extrapolation errors, or sparse sampling."

All the calculated uncertainties in the manuscript have been adjusted based on this updated analysis.

b) close to the emission source, due to impact of turbulence, persistent turbulent structures form that cause the cross-section mass in the plume to generate peak-to-trough structures that are advected downwind from the source (Galkowski et al.). As a thought experiment - if the speed of these structures in the studied cases was (unluckily) the same as advection speed of those structures, it might be that also the sampling of the plume at different distances was not independent, leading to potential biases in the estimations (worst-case: if the extreme peak (through) was always sampled – one would observe consistently positive (negative) bias in evaluated emission, respectively. The authors need to evaluate whether this synchronization of sampling and plume structure is responsible for the observed biases.

This is effectively describing the storage term, since the peak-to-trough structures shown in Galkowski et al. are due to build-up and release of emissions, cause by large-scale turbulent fluctuations in the advection speed. We hope that the additional explanation around storage (see response to comment 3 below) will help make this clearer. Storage is most likely related to the bias seen in Figures 3, 6, and 7 (Figures 3, 8, and 9 in the revised manuscript) and it is discussed extensively in the manuscript.

3. If my understanding is correct, what authors call "storage" is actually a momentary turbulent flux (positive or negative) – but it is resolved in the model, so not considered by author's definition of turbulent flux as described in Appendix B of Fathi et al. 2023. If my assumption is correct, then a more appropriate term here would be "large eddy turbulent flux". I would like to suggest adding discussion on relationship between "storage", turbulence and advective fluxes, as well as the effect of their interplay, somewhere in the study.

The following discussion is added in Section 2.3. *"A significant part of the storage term can be due to eddies and circulation at scales comparable to the control volume or flight time. As horizontal winds decrease (or increase), the total concentration within the volume will increase giving $S > 0$ (or decrease giving $S < 0$). Very generally, the horizontal turbulence term (E_{HT}) estimates flux due to boundary-layer turbulence, while the storage term (S) estimates flux due to mesoscale turbulence. However, as discussed in Fathi et al. (2021), storage can also include the effects of any non-steady-state conditions. For example, changes in atmospheric stability can modify the plume's buoyancy, moving the plume to different heights and resulting in changes in the horizontal advection speed of the plume."*

4. The description and analysis of the role of wind speed should be expanded. Only very rudimentary information about how the effective wind speed and direction was calculated is given (L182). How the wind was calculated for each screen (or group flight) is crucial, as the results are very sensitive to biases of U. Especially in Fig 7c and 7d, I have a strong suspicion that the wind speed and direction cause the sign shift in the bias, as the overall plume structure visible in Fig. 1 turns progressively to more southerly directions.. It might be that more accurate evaluation of wind direction could help reducing that bias – it stands to reason that in those areas far downwind the wind direction (and speed) is highly variable within the screen and assumption of a single-average wind is simply wrong. Authors might either test if local wind speed information can be interpolated (UAVs or aircraft usually carry wind sensors), another approach could be to detect the central plume path (see Kuhlmann et al 2020).

The following text is added in the 3rd paragraph of Section 2.3 (below Eq 3). "$U_{\perp}$ is the wind speed perpendicular to the screen at *each screen location* (s, z)*... Both* C *and* $U_{\perp}$ *are typically measured simultaneously (or close to it) during the flight, which accounts for variation in the wind pattern across the area of the screen.*".

While we understand that other techniques (such as satellite emissions analysis) typically use an average estimated wind speed, flight-based mass-balance analysis always uses winds measured along the flight path. Most publications of aircraft mass-balance approaches do not explicitly state this, but if this makes it clearer for a wider audience, we are happy to add this explanation.

5. Finally, I would like to point out that the results from the modelling of four simulations covering two afternoons, even after so detailed an analysis, is not sufficient to generalize the results. Statements that could be interpreted as general recommendations should be therefore avoided, e.g.: "*alone, a screen at a downwind distance of 4 km or more provides the same level of accuracy for the three types of sources investigated here (i.e. elevated stacks, small surface area sources, or a large surface area source)*". There is simply not enough proof to extrapolate these results to all cases, with local conditions (meteorological and otherwise) playing such a major role in the atmospheric transport in turbulent conditions. I therefore suggest to soften all such statements. The paper will not lose its (high) value, but transparency will be increased. I have marked some of such statements below.

We have modified all the text outlined in the comments below (as well as other instances) to emphasize that the results are for these specific atmospheric conditions and for these specific cases. We have tried to avoid generally extrapolating from these results.

Other comments:

L59: "*... and requires induvial plumes to be well defined and separate (e.g. Baray et al., 2018).*" – This makes sense if information on individual sources is required. If information on the cluster / group of sources is sufficient, there is no such need.

This was the point of the sentence – if there is a need for individual source information, footprint models can do that, but mass balance can only do that with sufficient separation of sources. To make this clearer, we modified the text as "*Estimating separate emission rates for each source* is more difficult to do with the mass-balance method and requires induvial plumes to be well defined and separate".

L80: "*This study aims to optimize...*" – here the authors indirectly imply that the results could also be extrapolated to dust particles – or at least this is how I understand it. While it might be true, it needs to clearly be stated in the study (also in the abstract, and in conclusions) that the tracers emitted in WRF are considered gaseous sources, and that typical dust processes like deposition etc. are not considered.

In the abstract we modified the text to "were investigated to determine emission rate retrieval accuracy *for emissions of a trace gas...*"

In Section 2.2 (end of 3rd paragraph), we add "*Emissions are all treated as trace gas. These results could be extrapolated to particulate emissions (which would be expected from an area source such as an open pit mine); however, dust processes such as gravitational settling and deposition are not considered here.*"

In the conclusions (near the end of the 2nd last paragraph) we add "*Although gravitational settling of particles or deposition (of gas or particles) to the surface could modify the concentration profiles, especially near the surface,* these results *generally* emphasize the need to constrain aircraft measurements…".

L95: *Case Studies and Locations* – perhaps "location"? The study is concentrated around Athabasca Oil Sands and facilities there.

Corrected.

L96: "*The model is run*" -> "The model is run in an LES mode…"

This part of the sentence was deleted (see point below).

L96: "dz ≈ 12 m -- I assume this is the height of the lowest layer - please state it clearly. Also, this information is given in sec. 2.2. again (with 11.2 m stated), please see my comment there.

We removed the resolution information and moved the sentence to the second paragraph of the model description.

L108-L117: I find this paragraph hard to read, consider revising. Possibly also moving to another section, since here the focus is on the extraction of data from the model, which doesn't fit the section title. Some suggestions follow:

We have removed this paragraph in its entirety. The overview is presented elsewhere (e.g. last paragraph of Section 1) and we agree that it doesn't fit in this section.

L108: "*In this study, we …* " – erase this sentence and fragment of the next until "To achieve this" – This is said again below, with higher information content.

Paragraph deleted.

L110: "*along flight paths similar to those conducted during*" – I think it's ok to use "same" or "matching" here.

Paragraph deleted.

L110: "*The super-resolution of our model-generated atmospheric fields allow us to sample data at temporal and spatial scales of airborne measurements without the need for interpolation of model generated fields.*" – some details important for study reproducibility are missing. How was the model sampled in horizontal and vertical? Was it simply using nearest-neighbour sampling? Or interpolation was used? If yes – were absolute heights used, or pressure, for vertical coordinate?

We added this information in Section 2.3 (5th paragraph). There, the text is added as "*The data (wind and concentration) along the flight path (x, y, z, t) within the model are sampled from the model values at the nearest grid location. No interpolation is done within the grid-cell or time-step. The*

sampling locations are then mapped to screen locations (s, z), and interpolation of the 2D screens is done with the kriging method…".

L121: a) neither T, p or c symbols are used later in the paper, consider dropping; b) please be specific, which moisture variable is archived? Relative humidity? Specific humidity?

a) Symbols are removed.

b) Changed to "*…water vapour mixing ratio*".

L124: "*~ 31 km*" – Please give 31.25 km exactly, this makes sense with 1:5 nesting ratio for WRF, approximation raises an eyebrow.

Changed.

L125: Was the vertical resolution forced to 11.2 m for all 40 grid levels? This is not a typical WRF configuration with hybrid model levels, so please state it clearly here. For comparisons against other modelling setups, it would also help to state how many vertical layers are present in the lowest 3km, please add this information here.

The grid spacing in the vertical is both refined and nested for the two finest domains of the model. Here we quote from the Fathi et al. (2023) Geosci. Model Dev. Paper that "*… $\Delta z = 11.62$ m for the first 40 full grid levels near the surface.*" For technical details of the model, it is better that the reader refers to that paper.

L127: Please state the spatial resolution for NARR data as well.

"*31.25 km resolution*" added.

L131: Please limit the description to sources and tracers relevant to this analysis.

Discussion of other sources not used in the analysis has been removed. The paragraph now begins with "*We use 7 modeled emission locations in this analysis, which are described in Fathi et al. (2023). The locations are shown in Figure 1. These are comprised of 4* elevated (stack) sources, two small area surface sources (surface mines), *and* a large area source (tailings pond).".

L136: 1. "in height" repeated 2. Please give exact heights of all four stacks 3. Please state their respective emissions – do they differ? Consider a table if they do. This is relevant for the analysis later.

1. Heights are added as "The existing stacks (CNRL1-4) *have respective heights of 114, 54, 30, and 54 m*".

2. We add "*Each of the 4 stacks emits at the same rate and the area sources all emit at the same rate per unit area.*".

L137: "*Each source emits a known amount*" -- 1. Is it meant that the emissions are known in real world, or prescribed in the model? Please make clear. Consider "Each source in the model emits a known amount" or "The emissions prescribed in the model E_s can be compared…"

Replaced with the former option.

L139: *"Here we evaluate three emissions scenarios: stacks"* -- "Scenarios" does not make sense in this context. "Emissions from group of emitters" are evaluated, consider this or similar.

We change this to "Here we *group the different emission source types together:* stacks (the sum of CNRL1, 2, 3, and 4), small area sources (the sum of MINE1 and MINE2), and the large area source (POND)*, and we investigate each of the three groups separately*."

L146: "*more than enough*" –> sufficient

Changed.

Figure 1, caption: "*All stacks are combined…*" -> It's the emissions that are combined. Consider: "Emissions from all stacks are followed using a single tracer in the model. Small dispersed area sources are grouped similarly". Also: degree symbol missing in coordinates.

Modified as suggested to "*Emissions from all stacks are followed using a single tracer in the model. Emissions from two small area sources are grouped similarly*."

Degree symbols added throughout text and in figure labels.

L165-184: I have my doubts about whether the full algorithm in this context, as most of the components except for the horizontal advective flux are immediately discarded. See my major comment 3.

See response to comment 3 above. A discussion is added around large eddy turbulent flux and the storage term.

L182: More details need to be given on how the wind was calculated. See major comment 4.

See the response to comment 4 above.

L186: "*The terms… must be ignored*" -- Wording. Actually they must not be ignored -- because that would mean we accept a presence of potentially large bias, as the mass escapes the volume. More precisely, it is reasonable to >assume they are negligible< - provided that there is no indication of mass on the higher levels of the flight, and no deep convection was observed –half a sentence that none of this was observed is worth adding.

The text "*must be ignored*" is modified to "*can be assumed negligible (provided there is no indication of mass on the higher levels of the flight, and no deep convection is observed).*"

L191: "*a 3-dimensional prism*" – please add "or a cylinder"

Added.

L201-202: When read first, it feels like contradicting L151. I suggest removing part of sentence "*representing a well-mixed concentration in the boundary-layer*" entirely.

We have deleted this part of the sentence.

L215: This discussion of the storage is very relevant to biases demonstrated later, but not highlighted in the discussion. If it's possible to evaluate the storage component in the previous

study numerically, why not use the same method to "correct" the emission estimates here for individual cases? See also my major comment 3 and comment to L581.

We have significantly expanded the manuscript to include this suggested analysis. Two new Sections have been added as follows:

"2.8 Storage Variability

As discussed in Section 2.3, Fathi et al. (2021) estimated the ratio of the storage term to the known emission rate (S/E_S) based on actual flight paths and Fathi et al. (2023) estimated S/E_S for a modeled flight path for different source emissions. Since the storage term is highly variable and the effect of large-scale turbulent fluctuations can change during the time it takes to fly a screen, we investigate the variability of the storage term for various flight lengths associated with the different flight configurations. The total integrated concentration within each control volume is calculated as a time series for the model run duration on each date. For each flight configuration (3 sources, 6 downwind distances), the control volume is defined as an area enclosed by the screen on the north-east side, extending south to a latitude 2 km south of the source and west to a longitude 2 km west of source, where the 2 km buffer accounts for any upwind diffusion from the source. The time-averaged storage is then determined as the average rate of change in integrated concentration within the volume over that period, which is positive for build up of emissions within the volume or negative for release of material from the volume. The period length investigated corresponds to the average flight time for a given source at a given distance. For example, the average flight length for the screen 6 km downwind of the small stack sources (on Aug 20, 16:20) is 908 seconds. The storage is then calculated (within the volume enclosing the small area source up to the screen at 6 km) for each 908 s period in the entire 137 min time series, and the standard deviation of these values (σ_S) is determined. While this cannot give us the exact value of the storage term for each flight investigated (since the plume is sampled at different points in time and space while the storage term is changing), this does provide a quantification of the relative uncertainty due to changing storage for different flight configurations on different dates. The resulting storage variability is discussed in Section 3.3."

"3.3 Storage

We calculate the variability in the storage term (σ_S) for the volume defined by each flight configuration (3 sources, 6 downwind distances), for the two flight dates, as discussed in Section 2.8. The two dates investigated use the flight configurations for the flight sets starting on Aug 20 at 16:20 (Fig. 1a) and the flight sets starting on Sep 2 at 16:20 (Fig. 1c). The resulting variabilities in S/E_S as a function of D (using the average flight length for that flight set) are shown in Figure 12. Generally, for the stack and small area sources, the variability in the storage term is minimum between $D = 6$ and 8 km, while the variability in the storage term for the large area source increases with D. For the stack and small area sources, the higher σ_S at small D is likely because these flights take less time (typically 1 to 2 min) which leads to a higher variability between flights since each flight is a snapshot of a changing large-scale flow field. This effect is more pronounced for the stack sources, relative to the small area sources, and the effect is not seen for the large area sources, since there would likely be more variability in a thinner wafting plume from stacks or small area sources compared to the spread-out plume associated with a large area source. The higher σ_S at

large D may be due to the larger volume enclosing the source and plumes, which encloses large-scale eddies and circulation, offsetting the reduced variability due to the longer flight durations.

The autocorrelation of the storage rate (S) time series (which is 147 min and 109 min long for the Aug 20 and Sep 2 dates respectively) gives a timescale of less than 3 min for all cases. Hence, we are 95% confident that storage term for any one flight will be within $\pm 2.03\sigma_S$ (35 degrees of freedom, JCGM, 2008). For the configurations investigated here, an optimal downwind distance of $D = 6$ or 8 km givens σ_S between 4.4 and 7% for Aug 20 and between 11 and 15% for Sep 2, suggesting an uncertainty as high as 14% and 30% respectively. This is in good agreement with the values of S/E_S reported in Fathi et al. (2021) (–3% for the Aug 20 flights and –29% for the Sep 2 flights) and those reported in Fathi et al. (2023) (up to 10.9% for Aug 20 and –27.5% for Sep 2). For large area sources, the uncertainty associated with storage can be reduced to 4% and 14% (for Aug 20 and Sep 2) by flying the screen closer to the source at $D = 2$ km."

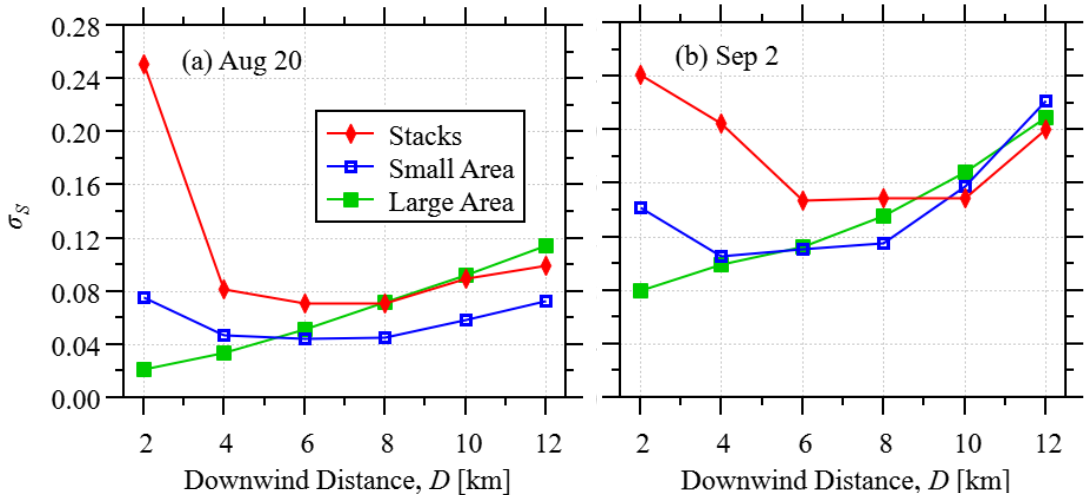


Figure 12. The variability (σ_S) in the storage term normalize by emission rate (S/E_S) as a function of the averaging period (τ) over the model run, from (a) Aug 20 16:20 to 18:47 and (b) Sep 2 16:20 to 18:09. The volumes are defined for the 3 source types (red line stacks, blue lines small area, green lines large area) and the 6 downwind distances ($D = 2$ to 12 km), which the furthest distances have the highest variability. For emphasis, the closest (2 km) and furthest (12 km) distances for each source type are highlighted.

L243: Please use another symbol. T is used for temperature or period of oscillation, both could theoretically be used in this study (e.g. period of circling around the source, where circular paths is discussed). In fact, temperature is also denoted as T in sec. 2.5.. To avoid confusion (especially in discussion), I strongly suggest simply Dh here (or similar).

We choose the variable ΔZ and have made replacements throughout. The delta demonstrates the differential spacing, which is distinct from the absolute distance measured by D, and the Z makes it clear that it is vertical.

L243: If T is set to 100m, then what's the point of optimizing it? Is that the base value? Please make clear

Modified to: "*...an initial value* of the vertical transect spacing... *is set to* $T = 100$ m,"

L245: Is 1 minute for turn a realistic time based on actual data from measurement campaigns?

Text added: "*(based on flight paths from Gordon et al., 2015 and Liggio et al., 2016)*".

L249: See major comment 1.

This is corrected and moved to Section 2.4.2.

L249: "*Based on our estimation...*" – sigma is simply a single measurement uncertainty estimate. Please erase or simplify.

As above.

L250: "*When comparing...*" – Is this relevant? Please clarify or erase.

The sentence is removed as it is no longer needed.

L255: In real world that would mean we have 10 instruments available. Perhaps add clarification whether this is meant to represent real-world situation where someone is flying 10 drones (unlikely for various reasons), or is just a method to estimate uncertainty. Related to major comment 2.

In the previous paragraph (where we first introduce that there are 10 flights in each set), we add the following text, "*For each set of 10 flights, each subsequent flight starts 1 minute later than the start of the previous flight. This offset is added to investigate the uncertainty in the estimated emission rate due to turbulent fluctuations with time scale on the order of 1 to 10 mins. Through the statistical analysis of multiple flights, we can also assess how effective repeated flights (or multiple sampling with 2 or 3 UAVs or aircraft) are in reducing the measurement uncertainty in the emission rate estimate.*".

L256: "*horizontal aircraft speed is randomly offset...*" Is this number according to real data? Based on my knowledge the variability of speed in UAVs in automatic mode is usually within 0.1 m/s at an altitudes up to 200 meters. When flown "manually" this value increases somehow (0.5 m/s – data from actual measurements) but having 3 m/s variability is unlikely, as these sort of conditions are not flight-permitting. For small aircraft change of wind speed by 3 m/s at higher altitudes is perhaps more likely, but then the momentum preservation law will prevents that to be >completely< random. And for larger aircraft this is simply impossible. Finally, the accumulation of the error is an entirely wrong assumption as either the automatic guidance systems, or the pilots will prevent "drifts" of the desired speeds and altitudes. This needs to be addressed, either by recalculating the procedure entirely, or by demonstrating that this does not lead to major biases in estimation.

After the "3 m/s" sentence, we add the text "*These random offsets, although potentially exaggerated compared to the variability of real flight speed or position, were found to produce visually similar flight paths compared to paths shown in Gordon et al. (2015). Given that this is a very subjective comparison, we investigate the effect of reduced offsets in Section 3.1.2 below. Although the analysis demonstrates that the effect of the randomized offset is small (<7% change in the average*

horizontal advective flux), the temporal and spatial offsets ensures that each of the 10 flights (for each D and ΔZ value) is distinct but generally sampling the same meteorological and emission conditions."

We have added a new Section with a new figure (which will be renumbered in the revised manuscript).

"3.1.2 Sensitivity to Random Offsets

As discussed in Section 2.4, at each 1-s timestep of the flight, the horizontal aircraft speed is randomly offset by a Gaussian random number with a standard deviation of 3 m/s, and the vertical position is offset by a Gaussian random number with a standard deviation of 1 m. To assess the sensitivity of the results to the scale of the offsets, we rerun the analysis for the set of flights on Aug 20 flight (at 16:20) at $D = 6$ km with both horizontal speed and vertical position offsets (3 m/s and 1 m respectively) simultaneously modified by a factor of 0 (i.e. no offset), 1/3, and 2/3. The resulting changes in estimated E_H/E_S and the variability (σ) are shown in Figure 5 (in addition to the 3 m/s and 1 m offsets used throughout the study). Using an evenly spaced, elevation-following grid with no offsets give a value of $E_H/E_S = 1.10$ with σ =28%. Adding a small amount of random offset (a factor of 1/3) to the grid increases E_H/E_S to 1.15. The offsets higher than that (factors of 2/3 and 1) both give $E_H/E_S = 1.17$. The difference in variability between flights within the flight set is < 1%. Hence, although there is a slight difference between no random offsets (even grid spacing) and the inclusion of random offsets (E_H/E_S of 1.10 versus 1.15), the results are not sensitive to the size of the offset over the range of values investigated here."

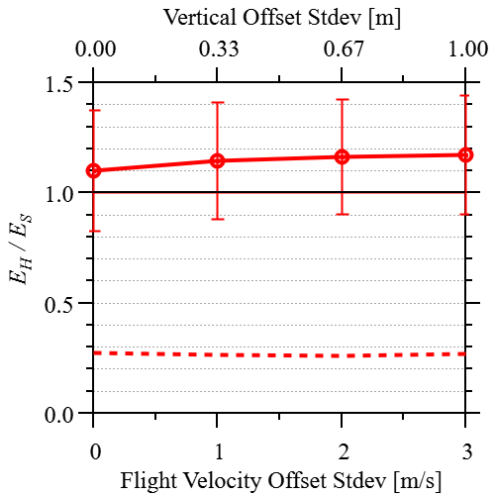


Figure 5. The variation in the ratio of horizontal advection flux (E_H) to the known emission rate (E_S) with change in the random offsets along the flight path for the Aug 20 flight set starting at 16:20. Error bars show one standard deviation (σ) calculated from 10 flights and the dashed lines show σ as absolute values.

L268: "*These screens are flown…*" - I think "flown" is confusing in this context - if the full screen are output at a single time, then perhaps it's better to use "sampled" here.

Changed.

L270: *We refer to these flights and the calculated emission rate values as "instantaneous".* ◊Linked to above; I suggest: "... to these calculated emission rates as instantaneous".

Changed.

L274: Here and throughout the text, I feel it would be beneficial to differentiate between a "single flight" and "10 subsequent flights". Consider "formation flight", "group flight" or even "echelon flight".

We prefer the term "flight set" or "set of 10 flights". This is added in Section 2.4 where the use of 10 flights for statistical investigation is first discussed, and the language throughout the manuscript is changed to refer to the "flight sets".

L281: "*turbulence and the and stability*" –repeated "and"

Corrected.

L283: "*model runs (using a criteria of 0.25 > Ri > –0.25 for neutral conditions)"* -- This is not a typical interpretation: please add a reference for range given if available. I've never encountered values below zero to be interpreted as neutral. Usually flows with Ri_b < 0.25 are treated as turbulent. See Stull, "Introduction to Boundary Layer Meteorology", Sec, 5.6.3. Fig 5.19., for example.

This was a mistaken interpretation. We have removed the text in the brackets and added the qualifier "...demonstrates that the conditions are *always turbulent and likely* unstable...".

L284: *"Temperature rises consistently during both afternoons, rising approximately*" – "rises, rising" - replace second with "by"

Changed.

L288: -- way >to< sample

Fixed.

L291: *"eliminated"* – should be "eliminating"

Fixed.

L305: "*calculation of .... means ...*" -> "calculating the screen length... results in a screen length that is..."

Modified.

L340: I think the critical point here is the temporal scale of the changes - these occur on high time frequencies, high enough to cause variability in estimated emissions between flights separated by 1 minute. The "storage" term here is a manifestation of the turbulent eddies transferring mass through the screen at highly variable rates. See my major comment 3.

We agree. We hope that the added discussion in response to Comment 3 has addressed this point.

Figure 3. a. This figure is only for stacks - and it should be noted in the caption. b. Red symbols are not mentioned - please add where appropriate. c. Please add Panel A/B/C/D references next to appropriate dates.

a) Added “for the stack sources”. b) The red symbols were mentioned in the 3rd line, but we started a new sentence with “The red circles…” to make that more clear. c) Panel identifiers are added next to the times in the figure caption.

L376: “The extrapolated concentrations…” - The way this is written it suggests (“average of…”) that more than 1 sampling was compared, but the text above says it was only the "first single instantaneous flight" was sampled and compared against the original screen. Please clarify if only one instance was compared, or the comparison was done to 10 flights).

This has been removed due to responses to the first reviewer’s comments about kriging analysis.

L378-380: The authors correctly spotted this effect for vertical motion but didn’t consider it for horizontal – see major comment 2.

As outlined in the response to Comment 2, we consider horizontal advection fluctuations in the storage term.

L393: “*Generally, flying…*” – I’m quite certain this is due to source being below 150 m and extrapolating without “seeing” most of the mass. It's quite clear from Fig 4, where at 2km the tracer concentration extrapolation < 150m underestimates concentrations in 14/16 cases (most of those quite clearly). Would require to look in detail on the model output over a longer period (if the output is available for several hours, then would be a good addition) to confirm without any doubt, but it’s quite logical - 2 km is not enough distance and time for the model to assure updrafts move the mass above 150m.

We agree with the reviewer, but it is not clear if any modification to the text is needed here.

L408: “*This transition from overestimation at small spacing to underestimation at larger spacing could be due to vertical movement of the plume opposite to the sampling direction, resulting in transects missing the plume centre at larger spacing.*” – Again, authors think only of vertical, but not of horizontal. See major comment 2.

As outlined in the response to Comment 2, we consider horizontal advection fluctuations in the storage term.

L413: sometime -> sometimes

Changed.

L435: “For the small area sources (Figs. 6a-d), the instantaneous flight horizontal…” – see comment for L393, same effect.

As above, it is unclear if modifications are required or if this just a general comment.

L445: “The relatively good agreement between instantaneous and non-instantaneous estimates implies that vertical motion of the plume does not result in over- or under-sampling.” It is also partially because for large source area the effective distance from the source is much larger – what

is given is calculated to the >edge< of a large source, so that the emission-centre point is much further upwind (Fig 1), and the effective signal is from areas well-mixed (far away) and not well-mixed (close to measurement). This deserves some expanded discussion as well, with "effective distance" or "distance to centerpoint" rather than distance to edge used for x if the comparison is to be fair.

The following text is added to Section 3.2.1 (at the end of the 2nd paragraph) "*Additionally, since D is defined as distance from the edge of the area source (as is necessary to sample the entire source area), emissions from the upwind side of the area source will have had more time to mix relative to the emissions from the downwind side of the are source. Hence, it would be expected that large area sources have smaller uncertainties for similar D values relative to small area sources.*".

L458: *"large area source would show substantially less uncertainty relative to a single flight sampling small area sources."* – delete "would", no need to hypothesise.

Changed.

L459: *"we expect"* – as above, "we estimate"

Changed.

L467: "*instantaneous area source flights*" -- instantaneous sampling maybe? See comment to L270.

Changed.

L477: "*however, this is…*" - Clearly something else negates this effect. Bias in wind speed or direction could be explained, especially since the model clearly predicts a large-scale change of wind direction, shifting to more southerly winds as the plume goes northwards. See major comment 4.

See response to Comment 4.

L484: "*For this source…*" – More precisely it should start with "for this source and these atmospheric conditions". See my major comment 5.

Text added.

L508: "*Scaling*" – this section doesn't have a corresponding entry in Methods. Reorganize, with expanded description of the method (and motivation for its use) moved to Section 2.

Text from Section 3.3 is moved to a new Section 2.9 (Scaling) in the Methods and it is rewritten to expand the description of the method and the motivation for its use.

L511: "*wind speed*" – horizontal, or also using W component?

Text "horizontal" added.

L511: "boundary layer heights are taken as…" –What was the method for PBLH evaluation here? State it clearly. Also, authors assume that PBLH did not change significantly – see comment below.

Following the comment below, PBLH is extracted form the model and this evaluation method is removed.

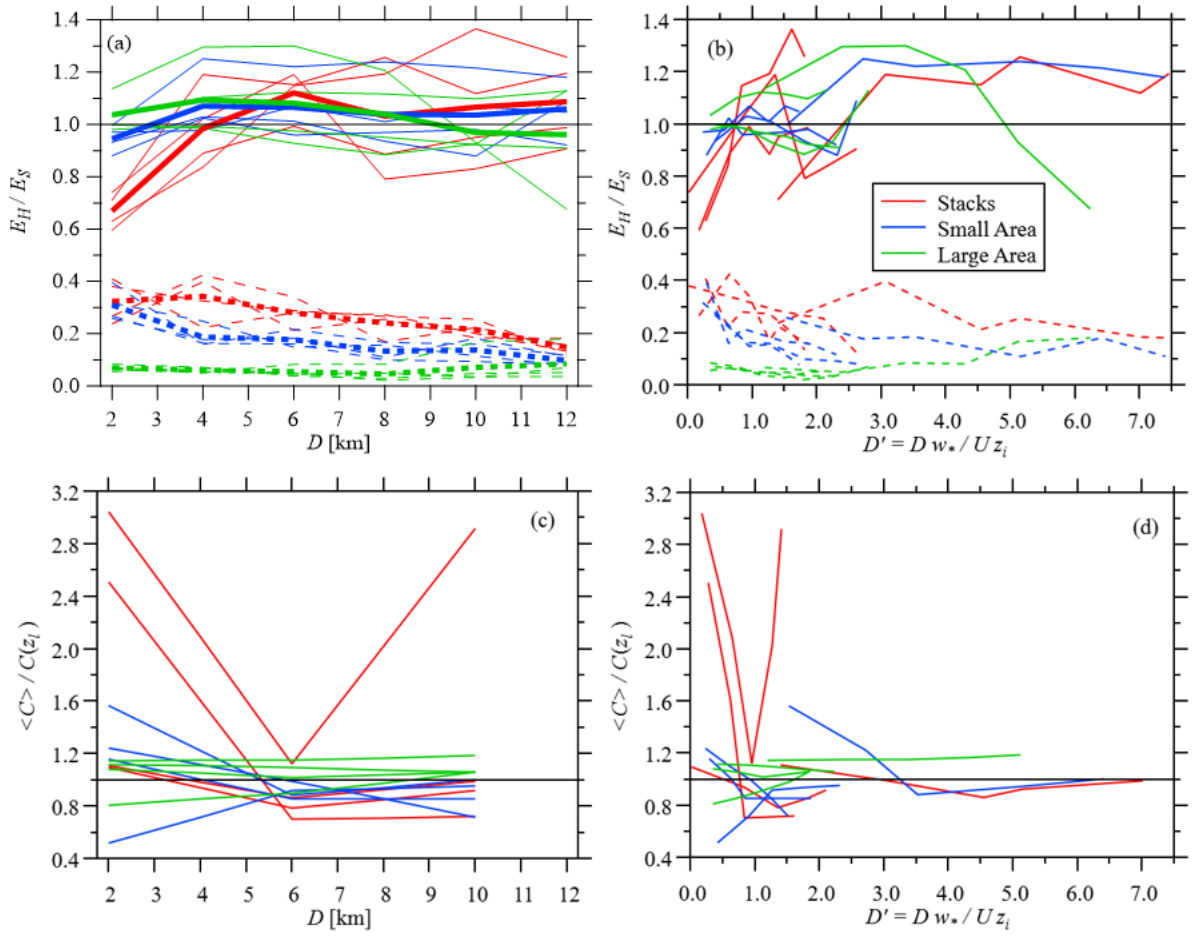
L516: "*The results are not collapsed…*" - Would it be better if actual PBLH was taken into the account, I wonder? High variability is possible - 16 UTC and 17UTC corresponds to approximately 9 and 10 local time in Alberta - PBLH development can be quite dynamic at this time, changes of 200 m per hour are typical for mid-latitudes in summer, so if the longest analysed time periods are 30 mins, then change of 100m is over 20% if the zi. I assume the PBLH field from the model is available, please give numbers here and discuss. Consider also the plume extent – single point values might not be representative.

We have extracted PBLH from the model. We add the following text to the new Section 2.9 (Scaling), "*The boundary-layer heights (z_i) are output from the model at the source locations. An average value of z_i is determined for each flight set and the effect of boundary-layer growth is discussed below.*"

In Section 3.3, we add the following "*Here we use an average z_i value for each flight set but note that this value can vary significantly during the flights. For all Aug 16:20 flight set, the model value of z_i is constant (824 m). It then grows linearly during the 17:20 flight sets, from 867 m to 1315 m (for the 40-min duration of longest flights at $D = 12$ km). During the Sep 2 flight sets, z_i increases from 528 m at 16:20 to 1078 m at 17:00 and then decreases from 1078 m at 17:10 to 952 m at 17:50. Hence, this normalization should be interpreted with some degree of caution. Although the variation is relatively small in most cases (more than 50% of the flight sets show less than 5% change in z_i during the flight durations), in some cases the increase in z_i can be up to 100%.*"

The modified Figure 13 (formerly Fig. 9) is shown below. Although this analysis has changed the figure slightly and introduces some uncertainty into the results, the interpretation of the results does not change significantly and very little of the following discussion is modified.

We also note that MDT is 6 hours behind UTC, so 16 UTC corresponds to 10 local time, although this doesn't change the reviewer's point.



Modified Figure 13 (formerly Fig. 9).

L527: "*for the Aug 20 17:20 stack flights at D= 10 km (see Fig. 4b), and it is unclear what would happen at further downwind distances for that flight.*" -- Large-scale change of wind direction is probably at play here and this breaks the method assumptions – see major comment 4.

Please see the response to comment 4. There is no assumed wind direction.

L545: "*Hence, based on the average estimate of E_H/E_S alone…*" – This reads as a general comment. I disagree that the evidence presented support this. See my major comment 5.

We have added the qualifying text "*for the cases studied here*".

L549: "*variability is seen instantaneous results*" – "seen in"

Corrected.

L555: "*Hence, 3 flights can be flown at D= 4 km in the same time it takes to fly one flight at D = 12 km. Taking the average of these 3 flights, reduces the uncertainty by a factor of 0.58 ($1/\sqrt{3}$). Hence,* …" – numbers flawed as based on wrong assumptions of statistics. See major comment 2. Also: "hence" is used twice.

We have corrected the double “hence” and have modified the statistics to account for n_{eff} as discussed above.

L561-567: Again, results are based on four eddy realizations. I find the sample size too small to derive such conclusions. See major comment 5.

We make the following edits here:

“...the results show that, for these cases, reducing the transect spacing...”

“For *the* area sources we investigate here, the variability...”

“For *the* small area sources *we investigate here*, increasing...”

“for the large area source *we investigate here*, increasing...”

L579: “*However, the results do demonstrate the potential to improve emission rate retrieval by accompanying any flight campaign with a strong modelling effort.*” - While I agree this statement is true in general, I do need to point out that this is not demonstrated in this study, as the results were not compared to actual measurement data here -- emission estimates were not "improved". Consider erasing. See also comment below (L581).

This is modified to “... the results *suggest that* emission rate retrieval *could potentially be improved* by accompanying any flight campaign with a strong modelling effort*, at least to help with understanding of the plume dynamics and behaviour*.”

L581: “*Reanalysis data combined with tracer release can be used to mimic flight actual patterns and estimate storage and release during actual flight time, thus reducing the most substantial uncertainty in the emission rate estimation.*” – If I understand the authors’ thought here, the model would require that to simulate exactly the same plumes, same eddies, as in reality. Do authors believe this is possible? The eddies are stochastic, and while can simulate realistic conditions, it's unlikely that we will reproduce exactly the same eddy pattern. And if we can't, then can the model help us correct estimations if we only have a single, or maybe two flights (as we often do?). Or does it only allow us to estimate uncertainty more realistically? Please comment.

This sentence is deleted.
