We want to thank John Miller for the review of our manuscript and the variety of helpful input and comments to improve this manuscript.

Review of Maier et al, 2023, submitted to ACP, by John Miller

“Potential of 14C-based versus ΔCO-based ΔffCO2 observations to estimate urban fossil fuel CO2 (ffCO2) emissions”

General comments:

This paper presents very promising results showing that continuous CO data, when ‘calibrated’ with discrete $D^{14}CO_2$ measurements to produce ‘pseudo-continuous’ fossil CO_2 mole fractions (and then averaged over a week to a month) has great potential in estimating urban fossil CO_2 emissions. To me, this is the most important result from the study and could be emphasized a bit more. I was impressed by the breadth of sensitivity tests that were conducted, which provide a lot of confidence in the results. The figures were clear (although I have a few small suggestions, and the writing was generally good; I have included some inline comments to help with clarity and English usage.

I am recommending 'accepted subject to minor revisions', because I don't think that at present any single suggestion I'm making is major, but in totality there are a lot of suggested/requested revisions. Additionally, as noted below and in the annotated .pdf I am interested to understand why formal random error was not estimated (or at least presented) in this study. I don't know if incorporating that information would constitute something 'major'.

The main objective of our study was to answer the question of which $\Delta ffCO_2$ information (discrete ^{14}C-based or continuous ΔCO-based $\Delta ffCO_2$) is best suited to estimate the seasonal cycle of fossil emissions in an urban region. To illustrate the information content of the $\Delta ffCO_2$ observations regarding the seasonal cycle of the $ffCO_2$ emissions, we switched off the seasonal cycle in the a-priori emissions. We could show that in our urban target region only the continuous ΔCO-based $\Delta ffCO_2$ observations (which were calibrated with ^{14}C-based $\Delta CO/\Delta ffCO_2$ ratios) lead to data-driven $ffCO_2$ emission estimates that are robust enough to be used to validate the seasonal cycles of the emission inventories. We demonstrate the robustness of our results by showing the a-posteriori seasonal cycles for several prior uncertainties (see the “spaghetti” plots in Fig. 3 and 4) and by performing additional sensitivity tests (see Fig. 5 and Fig. C1).

We want to emphasize that our study was not designed to *improve* the emission inventory within the main footprint of Heidelberg. We fully agree, that the calculation of the a-posteriori flux uncertainties would be essential for such a study. However, this would also require a careful estimation of the a-priori flux covariance matrix, thus detailed knowledge about the spatial and temporal correlations of the

prior emissions. However, we do not know these statistics for our regional target domain.

Therefore, we think that the shown spaghetti plots and sensitivity runs are better suited to demonstrate the robustness of the a-posteriori seasonal cycles and thus to answer our main research question than if a-posteriori flux uncertainties are presented, which would depend on a rather subjective choice of the a-priori flux uncertainties.

In addition to my overall positive impression of the paper, I list below numerous general and specific aspects that could (and in some cases need) to be improved. In no particular order, some general issues:

1. The Discussion section repeats much of what is said just above in the Results section. I personally prefer integrated results + discussion, because I think it is more efficient and (as is the case here) there are inevitably elements of 'discussion' in the results section. Sticking with current format is fine, of course, but the paper would be improved greatly by removing redundancy and focusing the discussion on new ideas and analysis. As just one example, I would be interested to learn more about what difference between Figs. 3 c and d tell us. On a related note, it would be interesting to see if you can derive some quantitative results from the large number of sensitivity results run, e.g., sensitivity of posterior to prior uncertainty and some estimate of posterior uncertainty given that this is not otherwise done.

We agree that our Discussion section had much redundancy. We tried to remove redundancy in the manuscript and merged the Discussion and Conclusions section, as was suggested by the other reviewer.

We performed the inversion runs in Fig. 3c and 3d to show that one hypothetical flask per week is not enough to get robust and plausible seasonal cycles in this urban region. The differences between Fig. 3c and 3d indicate that the inversion results are still determined by the model-data mismatch of individual (hypothetical) flasks. This is also illustrated by Fig. 1 (below), which shows the fits of the a-posteriori results to the observations. The inversion mainly reduces the largest model-data mismatches of individual winter flasks. We added this explanation in our manuscript (p. 13, l. 345ff).

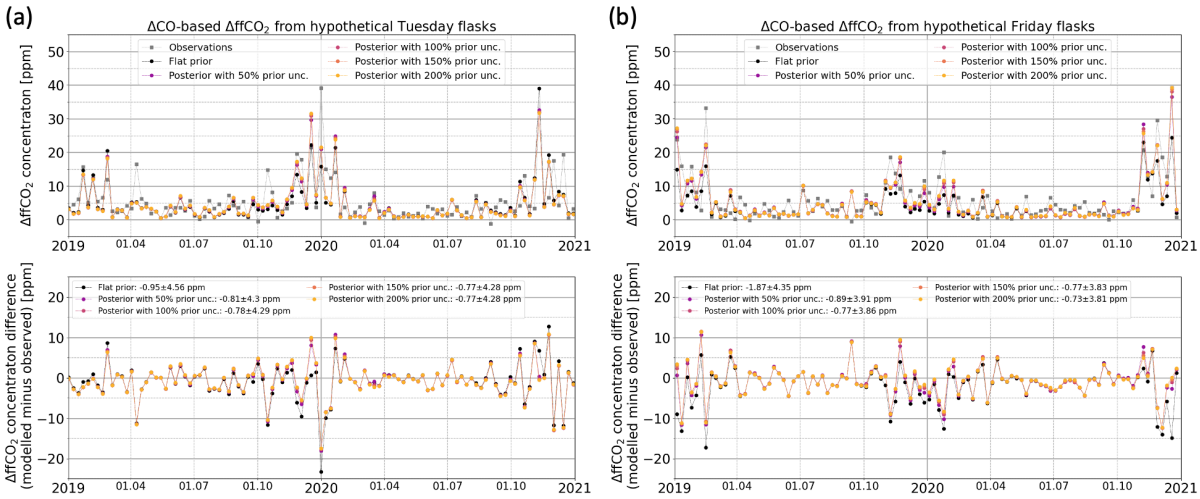

Figure 1. Fits of the flat a-priori and a-posteriori emissions to the observations for (a) hypothetical Tuesday and (b) Friday flasks. These fits correspond to Fig. 3c and 3d in the manuscript.

2. There is a lot of faith placed in the TNO inventory without having demonstrated (or discussed this). It implicitly serves as a truth metric. I would feel more comfortable with this if you discussed this explicitly saying that you do in fact treat TNO as a truth metric but also saying why it should be treated as such, especially with respect to its seasonality.

Additionally, there are numerous cases where the interpretation of results assumes the point source emissions to be perfectly accurate (all mismatch being assumed to result from transport uncertainty) and also the assumption that the spatial pattern of the area sources is perfectly known. These key assumptions need to be acknowledged more clearly, and the fact that they are not perfect assumptions needs to be recognized in the interpretation of results.

We agree, we placed a lot of faith in the TNO inventory by fixing the point source emissions and by optimizing only one spatial scaling factor for the area source emissions. We now point out potential inaccuracies in the TNO emission inventory (e.g. p. 14, l. 372ff; p.18 l. 495; p. 19 l. 534ff).

Our main purpose of showing the TNO seasonal cycles is to assess whether our a-posteriori seasonal cycles show a plausible amplitude and phasing. However, to get more trust into the TNO emissions, we now compare the TNO emissions in the nearfield of Heidelberg with emissions based on the EDGAR and GridFED inventory (see Fig. G1 in the revised manuscript). We also included a discussion about the differences between the inventories in our manuscript (p. 20, l. 557-569). While the EDGAR emissions are on average about 25% lower than the TNO emissions in the nearfield area of Heidelberg, the GridFED inventory shows ca. 23% larger emissions than TNO. Overall, the seasonal cycles of our top-down estimate are in the range covered by all three bottom-up inventories, thus inferring that we could indeed

reliably reconstruct the amplitude and the phasing of the seasonal cycle from flat a-priori area source $ffCO_2$ emissions with the ΔCO-based $\Delta ffCO_2$ observations.

3. In general, there is a need for more detail to be included in the paper. A thorough list is provided in line as comments in the marked up .pdf, but I'll mention some items here as well.

Thank you for this marked up .pdf. We respond to your comments in the .pdf directly. We tried to implement most of the suggestions.

a. I think it's important to quantify how different versions of the inverse model fit the observations. Reduced chi-squared, std. dev, and mean bias (please separate sd and bias instead of using RMSE which blends these) are important and easy to calculate metrics. These are especially important when trying to demonstrate things like overfitting. On a related note, it doesn't appear to be the case, but were any observations withheld for cross validation?

Thank you for these suggestions. We included fits of the a-posteriori results to the observations (see Fig. B1, Fig. C3 in the manuscript), and calculated mean biases and standard deviations (e.g. p. 12, l. 316ff; p.17, l. 450ff; p. 27, l. 768ff). For example, Fig. B1 illustrates the fitting of individual flasks with large model-data mismatches. For the ^{14}C-based $\Delta ffCO_2$ inversion, we also performed a reduced chi-squared analysis (p. 12, l. 320-324). However, the reduced chi-squared values range from 1.10 (for a prior uncertainty of 50%) to 0.97 (for a prior uncertainty of 150%), which already lead to large and unrealistic variabilities in the a-posteriori seasonal cycles. Therefore, we conclude that in our case the reduced chi-squared analysis might be less suitable to show overfitting.

b. A brief mention of how the non-fossil parts of the radiocarbon budget are treated in the construction of atmospheric CO2ff, especially the nuclear reactor flux of $^{14}CO_2$, would be useful.

We decided to not include this detailed information, as we extensively describe the construction of the ^{14}C- and ΔCO-based $\Delta ffCO_2$ concentrations and their uncertainties in the companion paper (Maier et al., 2023a, https://doi.org/10.5194/egusphere-2023-1237). We set up the companion paper with the intention to provide the observational basis for the present manuscript, which then can fully focus on the inversion.

c. The inversion methodology deserves some description, mainly basic aspects such as that it is not an 'analytical' inversion (i.e. an exact solution to the cost function minimum).

We included a more detailed description of the inversion methodology in the Appendix A in the manuscript (p. 23).

Given the small size of the state vector (the discretization of which would be good to explicitly describe) I would expect an analytical solution would be entirely possible just via a basic matrix inversion. Is there a reason this approach was not used and the more complicated R2005 approach was? The main significance is that a fully accurate posterior covariance could have been calculated allowing for presentation of analytically exact random error and also estimation of degrees of freedom, correlations over time, etc. And even in R2005 (according to my reading of it) a reasonable posterior covariance approximation should be available but none of these results are presented.

While the sensitivity tests address quite a few systematic error issues, it's unclear why the random errors were not presented.

We agree, the small size of the state vector might allow an analytical solution via matrix inversion. However, we decided to use the CarboScope framework instead due to two reasons: (1) The conjugate gradient algorithm used in the CarboScope framework leads also to a fast convergence because of the small state vector. (2) Using the CarboScope framework would allow the implementation of larger state spaces (e.g. to investigate in a next step the emissions from individual ffCO_2 sectors like heating or traffic without the need to set up a new inversion system). And this would avoid to set up two different inversion systems.

Regarding the a-posteriori random errors, please refer to our first answer above.

4. While a small point, I think it's important to clarify that the Delta(CO)-based method is actually based on both CO and 14C. Without this, readers may think CO and CO_2 alone have the capability to constrain fossil CO_2.

We fully agree, this may lead to confusion. We tried to make this clearer, also in the abstract (see p. 1, l. 21ff; p. 3, l. 77ff).

Specific comments are embedded as comments inline in the .pdf.

See our answers to your specific comments in the attached .pdf.