Referee#1

Yun et al have written an elegant and thought-provoking paper on error quantification for the OCO-2 MIP v10 dataset. The approach is a mixture of theory, where possible, and approximations, otherwise. It is certainly an interesting contribution, but its crude assumptions limit its scope to a scenario ("if we could assume that…, then we could conclude that…"). In fact, in practice, the main result seems to be the inference of... one element of the MIP protocol, namely the fact that fossil fuel fluxes were imposed on all inverse modeling systems. I can only encourage the authors to thoroughly revise their text in order to give it the right perspective.

[Response] We appreciate your constructive comments on this manuscript. We revised the manuscript to fully address your comments and suggestions. Detailed point-by-point responses to your comments and related revisions are presented below. The original comments are in black, and our responses are in blue color.

Detailed comment:

- Title: the flux errors here refer to the average of the flux ensemble, not to the individual flux sets. Please correct

  [Response] Based on the reviewer's suggestions, we changed the title as follow:
  "Quantification of regional terrestrial biosphere CO2 flux errors in v10 OCO-2 MIP **ensemble** using airborne measurements"

- 27: this main conclusion is not a finding but an input to the study (l. 118).

  [Response] The main finding of this study is that "actual errors in ensemble mean terrestrial flux estimates from v10 OCO-2 MIP are underestimated in regions with higher fossil fuel emissions compared to terrestrial biosphere fluxes." However, in the original main text, we did not provide information on the anthropogenic emissions in our analysis regions. In the revision, we added fossil fuel emission information to support our findings in the result section and Figure 6 as follow:

  "We find that the actual terrestrial biosphere flux errors are underestimated, particularly in regions where annual CO_2 emissions from fossil fuel combustion exceed annual terrestrial biosphere fluxes by 3-31 times. The airborne measurements carried out in mid-latitude North America, East Asia, and Southeast Asia are influenced by a broad region encompassing the United States, the eastern part of East Asia, and the western part of Southeast Asia where fossil fuel CO_2 emissions are 1,341, 2,443, and 815 Tg C year^{-1}, respectively. The first two regions are estimated as significant terrestrial biosphere CO_2 sinks, with estimated fluxes of −414 ± 279 (ensemble mean $\pm$ 1σ) and −561 ± 380 Tg C year^{-1}, in contrast to Southeast Asia (26 $\pm$ 118 Tg C year^{-1}). However, the CO_2 sinks are more than 3 and 4 times smaller than the fossil fuel CO_2 emissions, respectively. The recalculated terrestrial biosphere flux errors in these regions exceed the ensemble spread with values of 374, 643, and 211 Tg C year^{-1}. Observations in Europe and Australia, conducted over limited periods and specific locations, mainly represent certain areas in the western Europe and the southeastern part of Australia, where fossil fuel emissions (234 and 53 Tg C year^{-1}, respectively) are around four and five times greater than terrestrial biosphere sinks (−51 $\pm$ 34 and −10 $\pm$ 67 Tg C year^{-1}). The recalculated terrestrial biosphere flux errors in these regions are also larger than the ensemble spread, estimated at 65 and 114 Tg C year^{-1}, respectively. On the contrary, the most influential areas for the observation in Alaska and South America, encompassing the southeastern region of Alaska and the northern part of Brazil, characterized as a terrestrial biosphere sinks of −8 $\pm$ 11 Tg C year^{-1} and sources of 625 $\pm$ 387 Tg C year^{-1}, respectively, which are comparable to or more than 10 times greater than fossil fuel emissions (10 and 38 Tg C year^{-1}). The observation-based estimates of true terrestrial biosphere

flux errors are almost identical to the ensemble spread in both regions with values of 11 and 398 Tg C year^{-1}, respectively."

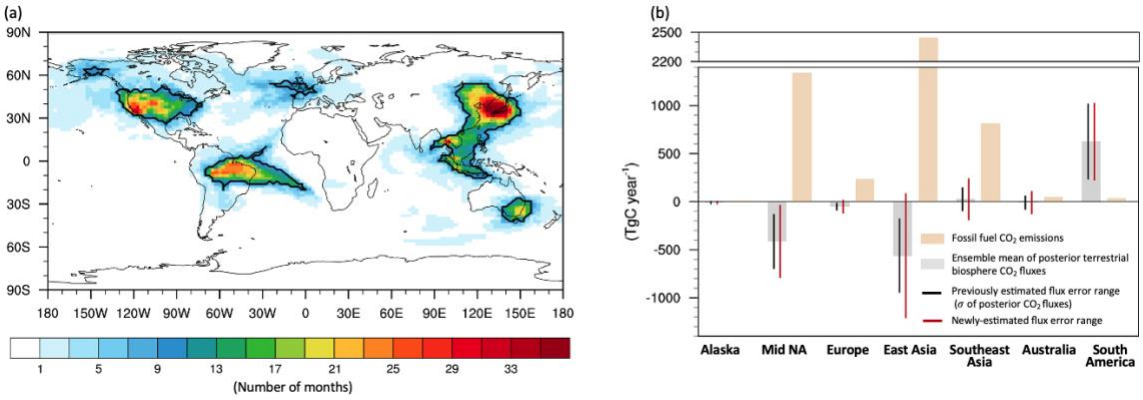

Figure 6: (a) Number of months selected as the effective area for airborne measurements. The outlined area represents selected areas for more than eight months or equal. (b) Annual total terrestrial biosphere CO_2 fluxes obtained from the ensemble mean of ten OCO-2 MIP models and annual total fossil fuel CO_2 emissions estimated from ODIAC data for each outlined area averaged over the period 2015–2017. The error bars in black and red indicate the one standard deviation of the inversion estimates and the newly estimated error range from this study, respectively.

Moreover, this study does not specify major reasons for the underestimation in true flux errors, but discusses possible causes for the underestimation. The one element of MIP protocol, "OCO-2 MIP models treat the fossil fuel emissions as true values and use the same dataset", could be one possible explanation for our findings but we did not (cannot) conclude that it is the main source of the underestimation in the true flux errors. This study aims to develop a framework to quantify the errors in regional terrestrial biosphere CO_2 fluxes estimated from an ensemble of inverse models. Evaluating impacts of each possible error source on ensemble flux estimates is out of the scope of this study. To better convey our main findings, we revised the corresponding sentences in the abstract as follows: "By identifying the most sensitive areas to airborne measurements through adjoint sensitivity analysis, we find that the underestimation of biosphere flux errors is prominent in eastern parts of Australia and East Asia, western parts of Europe and Southeast Asia, and midlatitude North America where the magnitudes of annual fossil fuel emissions exceed those of annual biosphere fluxes by 3-31 times over the three years. The regions with no underestimation are southeastern Alaska and northeastern South America where fossil fuel emissions are comparable to or less than biosphere fluxes." We also adjusted related statements in the main text.

- 53: RECCAP seems to be driven by GCP (https://www.globalcarbonproject.org/reccap/overview.htm) – what is the difference with the previous item (l. 52)?

  [Response] We agree that RECCAP activities utilizes the results from GCP. We revise that sentence as follow:
  "These projects include the TransCom project (Gurney et al., 2004; Houweling et al., 2015), which was first initiated in 1990s, as well as subsequent projects such as **the Global Carbon Project** (GCP; Friedlingstein et al., 2023; Ciais et al., 2022) and the Orbiting Carbon Observatory-2 (OCO-2) MIP (Crowell et al., 2019; Peiro et al., 2022; Byrne et al., 2023)."

- 66: please insert “explicitly” before “incorporate”, as the difference between systematic errors and error correlations can be subtle

  [Response] We revise that sentence as follow:

  “This Bayesian posterior uncertainty accounts for random errors in the prior fluxes and observations but does not **explicitly** incorporate systematic errors, thus providing a potential underestimate of the total posterior error.”

  77: it would be fairer to write that this method has no theoretical basis. “lacks” may suggest that there is hope to find one (or, please, elaborate).

  [Response] We revise that sentence as follow:
  “… but this method **does not have** an observational and theoretical basis and may not reflect actual errors”

- 93: the given definition of “error” is surprising, because “observed” is vague (by which technique?), and because the previous sentence is about fluxes.

  [Response] We revise that sentence as follow:
  “Here, "error" refers to the magnitude of the differences between **the true and estimated flux values**, without considering the sign.”

- 116: ten members only, covering only four transport models. How can their statistics be robust? Briefly touched in l. 483-7, but too late.

  [Response] Our estimates of flux and transport errors are computed from the ensemble spread of posterior CO_2 concentrations. Since the ensemble members encompass only four types of transport models, the estimated transport errors may not fully capture the actual transport errors. Thus, not only the disparity between the estimates and actual flux errors but also the discrepancy between the actual transport errors and their estimates could contribute to the differences between ERR_{TOT}^2 and $RMSE^2$. Therefore, we added the following discussions in the method and discussion sections in the revised manuscript, respectively:

  “However, our assumption regarding transport errors may be a strong assumption given that the transport errors are derived from 10 ensemble members, covering four different transport models, which might not fully capture the actual transport errors. We discuss how this assumption affects our key results in Section 4.”

  “We further investigate how our main results would be affected if the estimated transport errors deviate from actual errors by 20% and 40% of the difference between $RMSE^2$ and ERR_{TOT}^2. The ratio of regional mean of $h(err_{f_e})$ to $h(err_{f_t})$ increases by, on average, only up to 0.04 and 0.09 in the seven regions throughout the analysis period, respectively (Figure S11). In both cases, the estimated flux errors in mid-latitude North America, Europe, East Asia, and Southeast Asia still show significant underestimation at a 95% confidence level, while not in Alaska and South America. In Australia, characterized by a wide

uncertainty range, significant underestimation is also observed in the 20% cases, supporting the robustness of our findings. In the future OCO-2 MIP, the participation of inverse modeling groups using other transport models or meteorological forcing data might contribute to estimating transport errors closer to actual values."

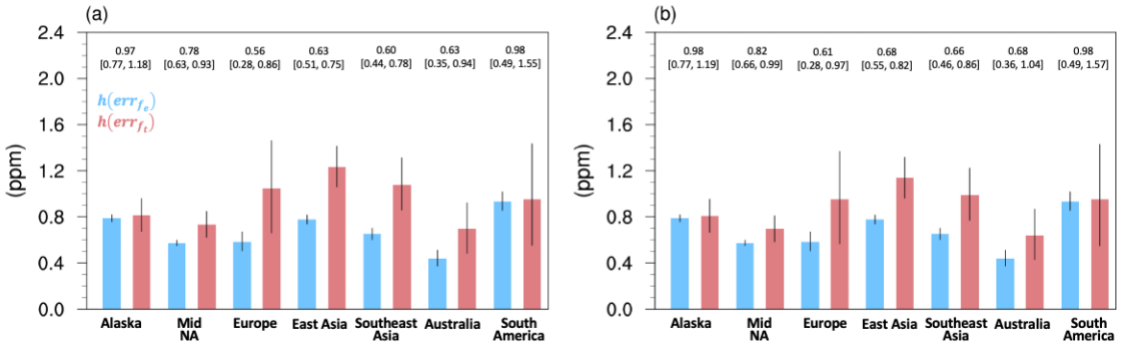

Figure S11. Mean values of monthly $h(err_{f_e})$, and $h(err_{f_t})$ for each region for the period 2015–2017 where (a) $h(err_{f_t})^2 - h(err_{f_e})^2 = 0.8 \cdot (RMSE^2 - ERR_{TOT}^2)$ and (b) $h(err_{f_t})^2 - h(err_{f_e})^2 = 0.6 \cdot (RMSE^2 - ERR_{TOT}^2)$. The numbers at the top of each panel indicate the ratio of the three-year mean $h(err_{f_e})$ to $h(err_{f_t})$. The error bars and error ranges represent the 95% confidence intervals derived from 1000 bootstrap samples of datasets.

- 148: please insert "below" before "to evaluate"

  [Response] We revise that sentence as follow:
  "We first employed the two matrixes defined in Eq. (1) and (2) **below** to evaluate ensemble posterior flux errors proposed by Liu et al. (2021)."

- 160: the way this exclusion is done biases the statistics towards the model values. Awkward.

  [Response] We agree on the reviewer's comment. We eliminated this process which excludes outliers and re-calculated all error quantities by using all observation data in the analysis. Since outliers comprised 0.05% of the total data, the newly computed results, particularly the ratio of ERR_{TOT} and RMSE and the ratio of $h(err_{f_e})$ to $h(err_{f_t})$, key metrics for assessing and quantifying regional terrestrial biosphere flux errors, do not exhibit significant differences compared to the previous results (Table R1).

Table R1 Mean values of the regionally averaged ratios of ERR_{TOT} to RMSE and the ratios of $h(err_{f_e})$ *to* $h(err_{f_t})$ for 2015–2017 with their 95% confidence intervals derived from 1000 bootstrap samples of datasets, calculated using atmospheric CO_2 datasets within the range of 1-5 km above ground level when excluding (previous results) or including outliers (revised results).

		Alaska	Mid NA	Europe	East Asia	Southeast Asia	Australia	South America
	ERR_{TOT}/RMSE	0.98 [0.89, 1.08]	0.91 [0.84, 0.97]	0.79 [0.61, 0.97]	0.87 [0.81, 0.94]	0.75 [0.65, 0.86]	0.73 [0.59, 0.87]	1.03 [0.83, 1.28]

Previous results	$h(err_{f_e})/h(err_{f_t})$	0.96 [0.76, 1.17]	0.75 [0.61, 0.90]	0.52 [0.28, 0.78]	0.64 [0.53, 0.77]	0.56 [0.41, 0.72]	0.59 [0.34, 0.87]	1.10 [0.51, 1.79]
Revised results	ERR_{TOT}/RMSE	**0.98 [0.89, 1.08]**	**0.90 [0.83, 0.97]**	**0.79 [0.61, 0.97]**	**0.84 [0.78, 0.91]**	**0.75 [0.65, 0.86]**	**0.73 [0.59, 0.87]**	**0.99 [0.79, 1.24]**
	$h(err_{f_e})/h(err_{f_t})$	**0.96 [0.76, 1.17]**	**0.74 [0.61, 0.88]**	**0.52 [0.27, 0.78]**	**0.59 [0.48, 0.70]**	**0.56 [0.41, 0.72]**	**0.59 [0.34, 0.87]**	**0.97 [0.49, 1.54]**

- 180-1: strong approximation. Also, the fact that the simulations were made at half-degree resolution, which is so much coarser than the measurements

[Response] Before addressing this comment, we'd like to mention that we changed the notation of representation error estimates from ERR_{o_r} to ERR_{REP} because in the revised version, we separately treat representation errors and observation (measurement) errors to convey our approach more clearly.

Our results show that the regional mean representation errors (ERR_{REP}) have lower monthly variability (i.e., standard deviation) ranging from 0.12 to 0.24 ppm compared to the variability of RMSE (0.24 to 0.45 ppm) and ERR_{MIP} (0.18 to 0.45 ppm) across all regions, except for South America, where the observational data are sparse (Figure 4). In addition, considering that we repeatedly used the high-resolution GEOS-Chem results for 2018 in the ERR_{REP} calculation for 2015-2017, we also examined the monthly variability of ERR_{REP} on an annual basis, focusing on the regions with year-round aircraft observation data. We found that their range is just 0.12-0.19 ppm and 0.13-0.20 ppm in North America and East Asia during 2015-2017, respectively, and 0.14-0.18 ppm over Southeast Asia during 2015-2016. Given that seasonal changes in atmospheric circulations and surface CO_2 fluxes are main drivers of the spatiotemporal variations of atmospheric CO_2 (Umezawa et al., 2018), the interannual variability of monthly CO2 variances within 2°x2.5° grid cells (VAR_{CO_2}) is anticipated to be lower than the sub-annual variability. Therefore, using the GEOS-Chem results from the same year repeatedly or from corresponding years for ERR_{REP} calculation would not lead to significant differences.

We revised the sentence by adding rationale for our assumption as follows:
"It is assumed that the variances do not vary significantly across years**, given relatively lower monthly variability of ERR_{REP} compared to that of RMSE and ERR_{MIP} (to be shown in Section 3.2).**"

Next, considering the disparity in spatial coverage between the GEOS-Chem model grid (0.5-degree) and observation points, we calculated the mean representation errors at a 1-degree grid scale and utilized it in our study. We agree on the importance of validating whether our representation errors, derived from simulated atmospheric CO_2 fields with 0.5-degree resolution, reasonably represents the actual spatial variability of CO_2 concentration within a 1-degree grid. For the validation, we used aircraft measurements from ACT-America project which provides extensive atmospheric CO_2 data across central and eastern North America spanning nine months for the period 2016-2019. Airborne observations do not provide simultaneous spatial distribution information of CO_2, unlike models. Thus, we calculated the variances of observed CO_2 concentrations within each 1°x1° grid cell with a 500 m vertical interval (from 1 to 5 km) at 3-hour intervals ($OBS\_ERR^2_{REP}$). On average, each grid box includes around 70 aircraft observation data. For comparison with the $OBS\_ERR^2_{REP}$, we sampled VAR_{CO_2}, derived from GEOS-Chem results, at the corresponding aircraft measurement times and locations at each grid box and computed their mean values ($MOD\_ERR^2_{REP}$). Monthly and regional mean OBS_ERR_{REP} and MOD_ERR_{REP} over North America show a significant positive correlation (r= 0.72, p <0.05) for the ACT-America project period (Figure S1). MOD_ERR_{REP} also has a similar mean value (0.62 [0.59, 0.64] ppm) with that of OBS_ERR_{REP} (0.49 [0.47, 0.51] ppm). This result supports that ERR_{REP}, based on 0.5-degree Geos Chem simulation results, could reasonably represent the actual mean observation representation error at a 1-degree grid scale.

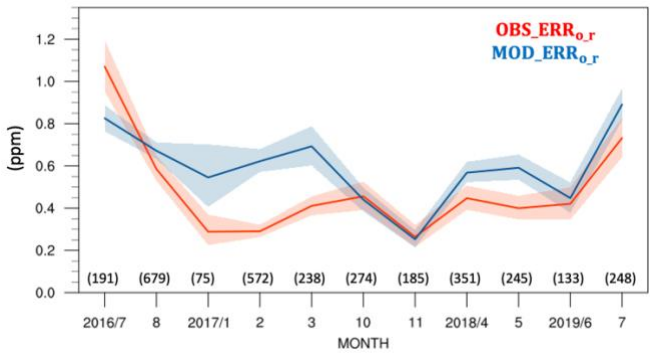

Figure S1 Monthly variations of OBS_ERR_{REP} and MOD_ERR_{REP} in the central and eastern North America for ACT-America project period during 2016–2019. The lines and shaded areas represent the mean and the 95% confidence intervals derived from 1000 bootstrap samples of the 1-degree grid-based monthly error quantities.

We added Figure S1 and above explanation on the supplementary information.

Umezawa, T., Matsueda, H., Sawa, Y., Niwa, Y., Machida, T., and Zhou, L.: Seasonal evaluation of tropospheric CO_2 over the Asia-Pacific region observed by the CONTRAIL commercial airliner measurements, Atmospheric Chemistry and Physics, 18, 14851–14866, https://doi.org/10.5194/acp-18-14851-2018, 2018.

197-198: This method mainly relies on this approximation, but I see no justification. You need to convince the reader that it is reasonable. Based, e.g., on Schuh et al (2019), "The research suggests that variability among transport models remains the largest source of uncertainty across global flux inversion systems", cited in l. 49, I would be surprised if it was, but please explain why I am wrong!

[Response] Our assumption is not in conflict with previous studies (Schuh et al., 2019) suggesting that transport errors are one of major sources of uncertainty in current global inversion estimates because transport errors are accounted for in both $RMSE^2$ and ERR^2_{TOT}. True transport errors are incorporated within $RMSE^2$ and estimated transport errors, computed from the ensemble spread among transport models used in OCO-2 MIP, are included in ERR^2_{TOT}. In this study, we assume that the estimated transport errors represent the actual transport errors. Thus, given that our estimates of representation errors reasonably depict actual representation errors, the difference between $RMSE^2$ and ERR^2_{TOT} mainly arises from the difference in the flux error variances ($\sigma^2_{f_t}$ and $\sigma^2_{f_e}$). However, we agree that our assumption regarding transport errors may be a strong assumption given that the transport errors are derived from 10 ensemble members, covering four different transport models, which might not fully capture the actual transport errors. We added discussions in the revised manuscript regarding the potential impact of errors in this assumption on our main findings as we addressed in our response to the reviewer's comment on L116. We also revised sentences to clarify the assumption applied in this study (i.e., lines 236-240 and 267-269).

- Figure 1, point 1): is actually about flux+transport errors

[Response] We agree with the reviewer's comment that the ratio of ERR^2_{TOT} to $RMSE^2$ indicates whether posterior flux and transport errors computed from the ensemble spread overestimate or underestimate

true errors in the ensemble mean of posterior fluxes and transport. However, as addressed in our response to the reviewer's previous comment on L197-198, this study assumes that the estimated transport errors from the ensemble spread among transport models used in OCO-2 MIP represent actual transport errors. To effectively convey our perspective, we added these sentences to the revised manuscript:

"**Given that ERR_{REP}^2 reasonably depict actual representation errors, $Ratio^2$ can indicate whether posterior flux and transport errors computed from the ensemble spread is an overestimation or underestimation of true flux and transport errors. In this study, we assume that the estimated transport errors from the ensemble spread among transport models used in OCO-2 MIP represent the true transport errors and the difference between $RMSE^2$ and ERR_{TOT}^2 mainly arises from the difference in the flux error variances ($\sigma_{f_t}^2$ and $\sigma_{f_e}^2$).** Thus, a ratio close to 1 indicates that the estimated posterior flux errors derived from the ensemble model spread are close to the true posterior flux error in the ensemble mean fluxes. A ratio greater than 1 means that the posterior flux errors are overestimated, and vice versa."

- 215: the concept of forward simulations obtained with a (backward-running) adjoint model is not intuitive. Did you use the adjoint to compute the Jacobian matrix and then did you run it forward?

  [Response] The GEOS-Chem Adjoint model integrates the forward GEOS-Chem and its derivative adjoint codes (https://wiki.seas.harvard.edu/geos-chem/index.php/GEOS-Chem_Adjoint_User%27s_Guide#Brief_overview). Within the model code, users have the option to choose whether to perform only forward simulation, in the same way used in a forward GEOS-Chem model, or to calculate adjoint sensitivity values as well. For this analysis, we choose the forward simulation options and derived simulated CO_2 concentration fields from the prescribed surface CO_2 fluxes.

  I have revised the sentence as follows to convey the information more clearly.

  "To get $h_{GC}(err_{f_e})^2$, we conduct a set of forward simulations using the GEOS-Chem transport model (within the GEOS-Chem Adjoint model v8.2j; Henze et al., 2007)."

- 220: if I understand it well, sub-monthly patterns are fixed, even though the comparison is to instantaneous measurements. The spread should be largely underestimated.

  [Response] In this study, our objective is to assess errors in the ensemble mean of posterior terrestrial biosphere CO_2 fluxes at regional scales on a monthly basis rather than a sub-monthly basis. Accordingly, all error quantities, including ERR_{TOT}, $RMSE$, $h(err_{f_e})$, $h(err_{f_t})$, err_{f_e}, and err_{f_t}, are computed on a monthly time scale. Thus, it is appropriate to use monthly posterior flux estimates for calculating atmospheric CO_2 errors solely attributed to the ensemble spread of monthly posterior fluxes from OCO-2 MIP ($h(err_{f_e})$).

- Eq. (13): my previous comments challenge it

  [Response] Please take a look at our responses to your previous comments.

- 226: again, I do not trust this hypothesis

  [Response] Please take a look at our response to your previous comment on L197-198.

  289: RMSE has already been defined

  [Response] We revised that sentence as follow:

  "…we compared RMSE with the sum of ERR_{MIP}, ERR_{REP}, and ERR_{OBS} (referred to as ERR_{TOT})."

- 300: Liu et al. (2022) is missing. I am looking for it to read the basis of "indicating most inverse models have common significant errors for this region".

  [Response] Apologies for any confusion. It's actually Liu et al. (2021), not Liu et al. (2022). We revised the citation information properly. Liu et al. (2021) showed an underestimation of posterior flux errors in CMS-Flux inverse model.

  Liu, J., Baskaran, L., Bowman, K., Schimel, D., Bloom, A. A., Parazoo, N. C., Oda, T., Carroll, D., Menemenlis, D., Joiner, J., Commane, R., Daube, B., Gatti, L. V., McKain, K., Miller, J., Stephens, B. B., Sweeney, C., and Wofsy, S.: Carbon Monitoring System Flux Net Biosphere Exchange 2020 (CMS-Flux NBE 2020), Earth System Science Data, 13, 299–330, https://doi.org/10.5194/essd-13-299-2021, 2021.

- 493: based on the above, I would challenge this statement.

  [Response] Please take a look at our responses to your previous comments.

Referee#2

Review.

This is an interesting and creative manuscript that I believe makes useful progress toward a challenging and important objective - evaluating uncertainty in the results of inverse estimates of biogenic CO2 fluxes. I believe the results and conclusions are justified given what I can gather from the data presented. My primary concern is the clarity of the text, both the methods and the results. At worst I could not understand some of the methods and results, and in other areas I think that I understand, but the presentation makes understanding a struggle.

I would encourage the authors to consider revising some of the presentation to make this important work more accessible. I have two main concerns.

[Response] We appreciate your constructive comments on this manuscript. We revised the manuscript to fully address your comments and suggestions. Detailed point-by-point responses to your comments and related revisions are presented below.

1. The authors work hard to explain the methods, but I struggled to follow. The figure is a good idea, and the appendix is very helpful. I found, however, that the terminology used, including the mathematical symbols used to define terms, was revealed gradually and irregularly. This makes reading the document difficult. I would strongly recommend presenting the most important variables and their definition up front and early in the text, and making sure to stick to that terminology and variable set throughout. I would make it easy for the reader to quickly look up the meaning of the most important variables used in the main results.

[Response] We appreciate your suggestions. We revised the method section to introduce main error statistics early on in the text. In addition, we revised many sentences to convey our approach more clearly. For detailed information about the revisions we made, please refer to our responses to your detailed comments.

2. Some of the presentation of results needs, in my opinion, to be rewritten. Some of the results are not organized into clearly written paragraphs, with a key finding as the topic sentence and discussion in the paragraph that explains the reasoning behind that key finding. Instead there are paragraphs that tend toward describing the figures, raising conclusions mid-paragraph or at the end of the paragraph, and those conclusions are clearly linked (in my mind) to the preceding text. I believe that rewriting some of the results and discussion (see detailed notes) will make the document easier to follow and more clearly illustrate what appear to be an interesting set of results derived from a creative set of methods.

[Response] Following your suggestions, we significantly revised many paragraphs in the results and discussion sections to ensure that our key findings and messages are effectively conveyed. For detailed information about the revisions we made, please refer to our responses to your detailed comments.

3. I have one question about the content. The number of airborne observations (which is not well defined, see my notes below) vary dramatically from region to region. I would expect this to have a much larger impact on the results than it appears to have. Heavily sampled regions (e.g. N America) don't appear much better understood than severely undersampled regions (e.g. S. America). Should we infer that intensive aircraft campaigns are not very beneficial, and that very limited sampling provides sufficient information for evaluating uncertainties in

inversions? Or that large investments in sampling does not greatly improve our understanding? Or is it safer to say that we have not yet learned how to use extensive data set to our greatest benefit?

[Response] The availability of airborne observation data in each region impacts both the reliability of our error statistics for quantifying regional flux errors and the area extents represented by these statistics (details described in our responses of your 21st and 34th comments). For example, the ratios of three-year mean $h(err_{f_e})$ to $h(err_{f_t})$, which are key metrics for quantifying regional flux errors (Figure 6h), have a smaller uncertainty range in the mid-latitude North America (0.75 [0.61, 0.89]; mean [95% confidence intervals]) and East Asia (0.59 [0.48, 0.70]), where wide and consistent data coverage are available, than other regions, particularly Europe (0.52 [0.27, 0.78]) and South America (0.97 [0.49, 1.54]), where observation coverage is sparse and intermittent. In addition, our three-year mean error statistics computed from data in mid-latitude North America and East Asia represent broad regions encompassing the United States and eastern parts of East Asia. In contrast, those computed from data in Alaska and Europe, where observation made for limited periods and at specific locations, represent much smaller regions. These results imply that intensive aircraft campaigns are critical for reliable evaluation and quantification of the errors in regional terrestrial flux estimates derived from inverse models. We believe that substantial investments in airborne sampling are undoubtedly beneficial in understanding the sources of errors in current inverse modeling and in estimating terrestrial biosphere CO_2 flux more accurately.

We included above explanations in the revised manuscript (lines 275-279, 448-450, and 513-527).

In sum I find the document very much worthy of publication, but in need of work on the presentation.

Detailed comments:

1. Lines 23 and 25. Are these references to fluxes specific to biogenic CO2 fluxes? At a few places in the abstract it isn't clear what fluxes are included. This gets especially confusing on line 27 when anthropogenic CO2 emissions are specified.

[Response] We revised these sentences as follow:

"…the observation-based error estimates exceed the atmospheric CO_2 errors computed from the ensemble spread of posterior **biosphere** CO_2 flux estimates by 1.33-1.93 times, …. By identifying the most sensitive areas to the airborne measurements through adjoint sensitivity analysis, we find that the underestimation of **biosphere** flux errors is prominent in eastern parts of Australia and East Asia, western parts of Europe and Southeast Asia, and midlatitude North America where the magnitudes of annual fossil fuel emissions exceed those of annual biosphere fluxes **by 3-31 times over the three years**. The regions with no underestimation were southeastern Alaska and northeastern South America **where fossil fuel emissions are comparable to or less than biosphere fluxes.**"

2. Line 36-37. English needs some work.

[Response] We revised these sentences as follow:

"Accurate estimates of regional terrestrial biosphere carbon fluxes and their uncertainties are, therefore, crucial for monitoring changes in terrestrial carbon sinks."

3. Line 40. final phrase is left dangling.

[Response] We revised these sentences as follow:

"Atmospheric CO_2 inverse modeling is one of the widely employed approaches to estimate terrestrial and air-sea CO_2 fluxes **by assimilating observed atmospheric CO_2 concentrations**."

4. Line 48. I'm not sure what "systematic errors in …. inversion setups" means.

[Response] We revised these sentences as follow:

"However, concerns have been raised that the inverse modeling results are **sensitive to the selection of transport models, prior flux datasets, and data assimilation techniques** that are not accounted for in the Bayesian framework."

5. Lines 85-90. This is tough to follow. But let me try the methods, then perhaps this will be clearer.

[Response] We revised these sentences as follow:

"We quantify the errors in ensemble mean estimates of posterior atmospheric CO_2 by comparing them with the airborne CO_2 data. We then estimate the contributions of various error components (e.g., representation, observation, transport, and flux errors) to the observation-model difference in atmospheric CO_2 and isolate the contribution of terrestrial flux errors. Next, we identify the areas that these airborne CO_2 are most sensitive to and quantify the annual biosphere flux errors in these areas."

6. Lines 91-92. If the objective focuses on the use of airborne observations, it might help to include some description of these observations and their suitability for this task in the introduction.

[Response] We appreciate the reviewer's suggestion. We include the following sentences in the introduction in the revised manuscript:

"This study uses more than 833,000 airborne CO2 measurement data collected at 1-5 km altitude from 20 different measurement projects (e.g., Baier et al., 2021; Miller et al., 2021; NOAA Carbon Cycle Group ObsPack Team, 2018; Schuldt et al., 2021a; 2021b). These data have broader spatial coverage and are less influenced by local sources compared to surface CO_2 data, thus capturing signals from regional surface CO_2 fluxes."

7. Line 96. I'm puzzled by the statement, "an approximation of RMSE." Maybe, "RMSE in the elements of the ensemble"? I'm not sure that is clearer.

[Response] The previous expression was not clear, so we revised these sentences as follows:
"First, we define two quantities: 1) the root mean square errors (*RMSE*) between the ensemble mean of posterior CO_2 concentrations and observed CO_2 concentrations, and 2) ERR_{TOT} (Section 2.3). $RMSE^2$ represents the true errors in OCO-2 MIP ensemble mean of CO_2 concentrations including representation errors (σ_r^2), observation errors (σ_o^2), true flux errors projected onto CO_2 concentration ($\sigma_{f_t}^2$), transport errors (σ_t^2), and error covariances between the preceding two terms ($cov(\sigma_{f_t}, \sigma_t)$). ERR_{TOT}^2 is the sum of the estimated error components, defined as the sum of ERR_{REP}^2, ERR_{OBS}^2 and ERR_{MIP}^2. ERR_{REP}^2 and ERR_{OBS}^2 indicate representation errors (σ_r^2) and observation errors (σ_o^2), respectively. ERR_{MIP}^2 is the sum of estimated flux errors projected onto CO_2 space ($\sigma_{f_e}^2$) and transport errors (σ_t^2), and their error covariances ($cov(\sigma_{f_e}, \sigma_t)$), computed from an ensemble spread of posterior CO_2 concentrations."

8. Line 99. Next? Did you just present this as (2) in line 96?

[Response] In previous manuscript, the analysis described in line 99-100 differs from that described in (2) in line 96. In line 96, we defined ERR_{TOT}^2 (=$ERR_{REP}^2 + ERR_{OBS}^2 + ERR_{MIP}^2$), where ERR_{MIP}^2 indicates the sum of estimated flux errors projected onto CO_2 space and transport errors, and their error covariances. ERR_{MIP}^2 is computed from an ensemble spread of "posterior CO2 concentrations". However, in the analysis described in line 99-100, we derived atmospheric CO_2 errors due to only the ensemble spread of "posterior CO2 flux estimates" ($h(err_{f_e})$) through transport model simulations. We revised these sentences to convey our approach more clearly as follow:

" … ERR_{MIP}^2 is the sum of estimated flux errors projected onto CO_2 space ($\sigma_{f_e}^2$) and transport errors (σ_t^2), and their error covariances ($cov(\sigma_{f_e}, \sigma_t)$), computed from an ensemble spread of posterior CO_2 concentrations. … Next, we calculate the estimated flux errors projected onto atmospheric CO_2 ($h(err_{f_e})$) through atmospheric transport simulations (Section 2.4)."

9. Line 101. What are the true errors? Does this differ from the ratio exercise described earlier?

[Response] RMSE represents the true errors in OCO-2 MIP ensemble mean of CO_2 concentrations including true flux errors projected onto CO_2 concentration and transport errors, error covariances between the preceding two terms, representation errors, and observation errors. In line 101 of the previous manuscript, the true errors indicate the true errors in ensemble mean posterior fluxes projected onto CO_2 space that are included in the RMSE.

In addition, earlier analysis based on the ratio between ERR_{TOT} and RMSE is for evaluating whether the flux errors computed from ensemble spread of posterior flux estimates overestimate or underestimate the true errors in ensemble mean of flux estimates of OCO-2 MIP models. However, in the later analysis, we quantify the true flux errors projected onto CO_2 space by isolating them within the RMSE.

We revise that paragraph to convey our approach more clearly (lines 96–114 in the revised manuscript; or please refer to the response in the following comment)

10. Figure 1. This is a nice idea, but the terms in this figure need to be defined. At present these terms don't match the terms in the text, and there are many undefined terms in the figure.

[Response] We appreciate the reviewer's suggestion. We revise that paragraph to match with the terms included in Figure 1. In addition, we changed the notation of representation errors from $\mathrm{ERR}_{\mathrm{o\_r}}$ to $\mathrm{ERR}_{\mathrm{REP}}$ and included this term in Figure 1. as follow:

"First, we define two quantities: 1) the root mean square errors (*RMSE*) between the ensemble mean of posterior CO_2 concentrations and observed CO_2 concentrations, and 2) ERR_{TOT} (Section 2.3). $RMSE^2$ represents the true errors in OCO-2 MIP ensemble mean of CO_2 concentrations including representation errors (σ_r^2), observation errors (σ_o^2), true flux errors projected onto CO_2 concentration ($\sigma_{f_t}^2$), transport errors (σ_t^2), and error covariances between the preceding two terms ($cov(\sigma_{f_t,}\sigma_t)$). ERR_{TOT}^2 is the sum of the estimated error components, defined as the sum of ERR_{REP}^2, ERR_{OBS}^2 and ERR_{MIP}^2. ERR_{REP}^2 and ERR_{OBS}^2 indicate representation errors (σ_r^2) and observation errors (σ_o^2), respectively. ERR_{MIP}^2 is the sum of estimated flux errors projected onto CO_2 space ($\sigma_{f_e}^2$) and transport errors (σ_t^2), and their error covariances ($cov(\sigma_{f_e,}\sigma_t)$), computed from an ensemble spread of posterior CO_2 concentrations. Here we separate representation errors from transport errors for computational purpose. The ratio between ERR_{TOT} and *RMSE* is then used to evaluate whether the estimated flux errors, computed from the ensemble spread of posterior fluxes, overestimate or underestimate the true errors in the ensemble mean fluxes. Next, we calculate the estimated flux errors projected onto atmospheric CO_2 ($h(err_{f_e})$) through atmospheric transport simulations (Section 2.4). With $h(err_{f_e})$, ERR_{TOT}, and *RMSE*, we derive the true errors in ensemble mean of posterior fluxes projected onto CO_2 space ($h(err_{f_t})$). Then, we identify the areas where these airborne observations are most sensitive to using an adjoint sensitivity analysis and calculate the estimated posterior flux errors over these regions (err_{f_e}). Assuming a linear observation operator, the study finally computes the true errors of the ensemble mean posterior fluxes over the identified sensitive areas (err_{f_t}) by applying the ratio between $h(err_{f_t})$ and $h(err_{f_e})$ to err_{f_e}."

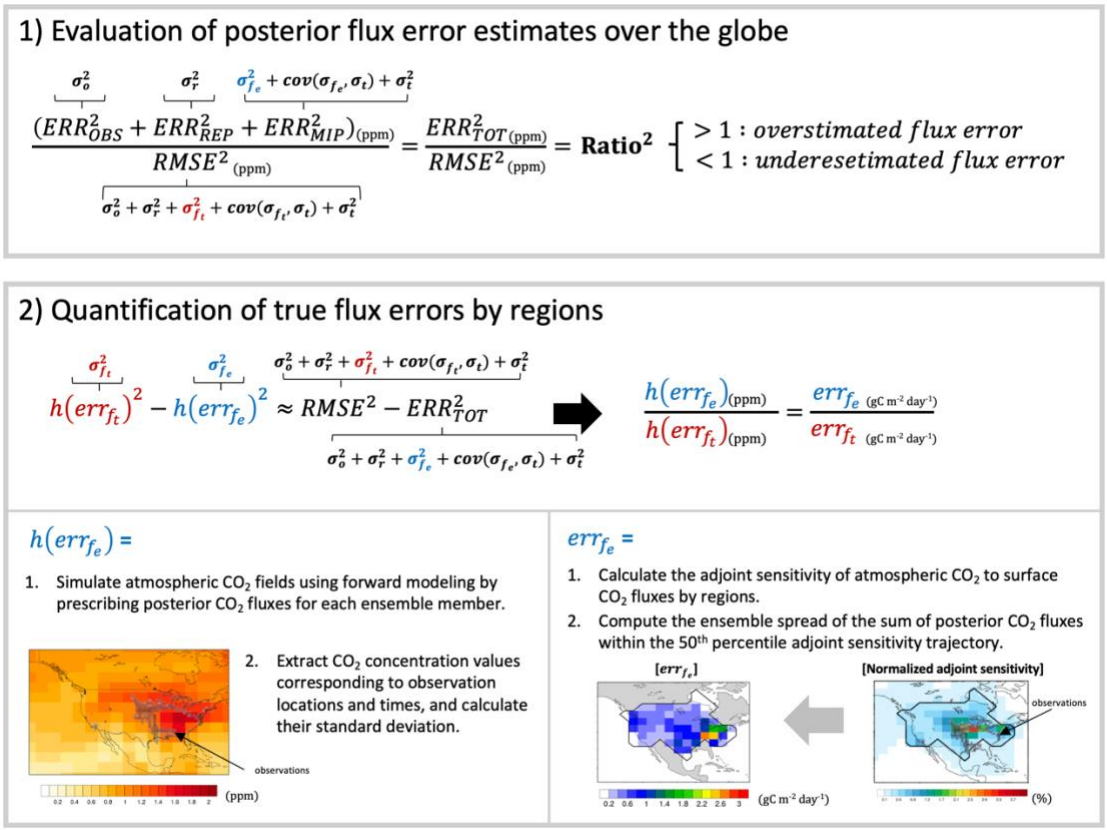

Figure 1: Flow chart summarizing the process of evaluating and quantifying errors in ensemble mean of regional posterior fluxes. RMSE² is the mean square errors between the ensemble mean of posterior CO_2 concentrations and observed CO_2 concentrations. ERR_{REP}^2 and ERR_{OBS}^2 denote estimates of observation errors and representation errors, respectively. ERR_{MIP}^2 is an ensemble spread of posterior CO_2 concentrations. ERR_{TOT}^2 is defined as the sum of ERR_{REP}^2, ERR_{OBS}^2, and ERR_{MIP}^2. err_{f_e} and err_{f_t} are estimates of flux errors, defined as an ensemble spread of posterior fluxes, and their true values. $h(err_{f_e})$ and $h(err_{f_t})$ are estimates of flux errors projected onto CO_2 concentrations and their true values. σ_o^2, σ_r^2, $\sigma_{f_t}^2$ ($\sigma_{f_e}^2$), σ_t^2, and cov(σ_{f_t},σ_t) indicate the types of errors represented by the error statics, namely observation errors, representation errors, true (estimated) flux errors projected onto CO_2 concentration, transport errors, and error covariances between the preceding two terms, respectively.

11. Line 126. I would recommend adding citations that document these field campaigns.

[Response] We include citation information for the measurement campaigns in both the text and Table 1 of the revised manuscript.

"The dataset includes two airborne measurement campaigns (Atmospheric Tomography Mission; ATom; **Thompson et al. 2022** and O_2/N_2 Ratio and CO_2 Airborne Southern Ocean Study; ORCAS; **Stephens et al. 2018**) over the ocean, as well as 18 campaigns over land."

12. Figure 2. The caption refers to the number of airborne measurements. What constitutes one airborne measurement? Many aircraft campaigns have continuous observations and gigabytes of data. Please explain the quantization of the data that is used in this figure. If this number of the number of 1x1 degree grids with an

observation, what is the temporal unit for an observation? If the same location is measured for 100 hours over 10 days within one month, is that one measurement or ten or 100 measurements?

[Response] Figure 2a illustrates the total number of airborne measurements data used in our analysis. Figure 2b shows the number of 1°x1° grid-points where more than 10 observations were made within each region for every month. If observation is made at the same location (i.e., same grid point) 100 hours over 10 days within one month, it is considered as one grid point. We revise the caption of Figure 2 to convey their definition more clearly:

"Figure 2: (a) Total number of airborne measurement data used in this study at each 1°×1° grid point and (b) the number of 1°×1° grid-points, where more than 10 data is available, within each region and each month for the period 2015–2017."

13. Table 1. Please include citations for data sets when possible. I am sure, for example, that there is a data citation available for AToM observations.

[Response] We include citation information for all data sets in Table 1 of the revised manuscript.

14. Line 149. "simulated atmospheric CO2 mole fractions"? And please explain, "the observed one". What is "the observed one"?

[Response] We revised these sentences to convey our approach more clearly as follow:
"One is RMSE between the ensemble mean of posterior atmospheric CO_2 from OCO-2 MIP models and the atmospheric CO_2 from airborne measurements, …"

15. Line 153. "the 1x1 grid cell"

[Response] We revised it as follow:
"… within **each** 1°×1° grid-cell in each month …"

16. Line 153. What constitutes one airborne measurement? The continuous aircraft campaigns have MANY more measurements than is suggested by Figure 2. Please explain your definition of one measurement.

[Response] In this study, all error statistics such as RMSE are computed using airborne measurement data made within each 1°x1° grid-cell during a month. In other words, all airborne measurement data recorded within a single grid point for one month provide one measurement information for evaluating inversion estimates. We revised that sentence in the revised manuscript:

"$\overline{h_i(\hat{x})}$ is the ensemble mean of posterior atmospheric CO_2 sampled at the time and location of the i^{th} airborne observation $y_{o,i}$, within each 1°×1° grid-cell in each month. *N* is the monthly total number of sampled data at each grid-cell. *M* is the number of ensemble members (i.e., 10). A single monthly *RMSE* value is computed using *N* measurement data at each grid-cell. The number of *RMSE* values is calculated per month within each region corresponds to the number of grid-cells shown in Figure 2b."

17. Line 159-160. I don't believe that the ensemble mean accounts for transport errors. The ensemble includes them, at least as represented by the ensemble.

[Response] We agree with the reviewer's comment. Transport errors can be estimated from the ensemble spread of inverse model estimates, rather than from the ensemble mean values. We revised the sentence and relocated it to the part where ERR^2_{MIP} (ensemble spread of posterior CO2 concentrations) is introduced (line 194–197 in the revised manuscript).

"Different from Liu et al. (2021) which used only one transport model, ERR^2_{MIP} accounts transport errors because posterior atmospheric CO_2 were generated by multiple types of transport models in OCO-2 MIP driven by different meteorology fields. Thus, ERR^2_{MIP} term accounts for transport errors, but not representation errors due to the coarse spatial resolution of these transport models with the highest spatial resolution being 2°×2.5°."

18. Line 161. I don't object to removing these outliers, but I'm not sure this ensures robust error estimates.

[Response] We eliminated this process which excludes outliers and re-calculated all error quantities by using all observation data in the analysis. Since outliers comprised 0.05% of the total data, the newly computed results, particularly the ratio of three-year mean ERR_{TOT} and RMSE and the ratio of three-year mean $h(err_{f_e})$ to $h(err_{f_t})$, key metrics for assessing and quantifying regional terrestrial biosphere flux errors, do not exhibit significant differences compared to the previous results (Table R1).

Table R2 Mean values of the regionally averaged ratios of ERR_{TOT} to RMSE and the ratios of $\boldsymbol{h(err_{f_e})}$ to $\boldsymbol{h(err_{f_t})}$ for 2015–2017 with their 95% confidence intervals derived from 1000 bootstrap samples of datasets, calculated using atmospheric CO_2 datasets within the range of 1-5 km above ground level when excluding (previous results) or including outliers (revised results).

		Alaska	Mid NA	Europe	East Asia	Southeast Asia	Australia	South America
Previous results	ERR_{TOT}/RMSE	0.98 [0.89, 1.08]	0.91 [0.84, 0.97]	0.79 [0.61, 0.97]	0.87 [0.81, 0.94]	0.75 [0.65, 0.86]	0.73 [0.59, 0.87]	1.03 [0.83, 1.28]
	$h(err_{f_e})/h(err_{f_t})$	0.96 [0.76, 1.17]	0.75 [0.61, 0.90]	0.52 [0.28, 0.78]	0.64 [0.53, 0.77]	0.56 [0.41, 0.72]	0.59 [0.34, 0.87]	1.10 [0.51, 1.79]
Revised results	ERR_{TOT}/RMSE	**0.98 [0.89, 1.08]**	**0.90 [0.83, 0.97]**	**0.79 [0.61, 0.97]**	**0.84 [0.78, 0.91]**	**0.75 [0.65, 0.86]**	**0.73 [0.59, 0.87]**	**0.99 [0.79, 1.24]**
	$h(err_{f_e})/h(err_{f_t})$	**0.96 [0.76, 1.17]**	**0.74 [0.61, 0.88]**	**0.52 [0.27, 0.78]**	**0.59 [0.48, 0.70]**	**0.56 [0.41, 0.72]**	**0.59 [0.34, 0.87]**	**0.97 [0.49, 1.54]**

19. Line 175-180. I may just be tired, but I am having a very hard time following this discussion. This is an interesting approach to evaluating uncertainty. It would be great if this could be explained more clearly. Figure 1 is an interesting complement to this text, but it isn't cited at all in this text. Perhaps you could clarify your methods by connecting the terms in Figure 1 explicitly to this text and to Appendix A.

[Response] We appreciate the reviewer's suggestion. In the revised manuscript, we changed the notation of representation errors from ERR_{o_r} to ERR_{REP} and included this term in Figure 1. We also revised sentences to clearly convey our approach for estimating representation errors:

"To obtain representation errors and observation errors not captured by ERR^2_{MIP}, we additionally calculate ERR^2_{REP} and ERR^2_{obs}, respectively. ERR^2_{REP} indicates the representation errors (σ_r^2) in *RMSE*^{2} as shown in Fig. 1 and is defined as a spatial variability of atmospheric CO_2 within a 2°×2.5° grid cell written as:

$$ERR^2_{REP} = \frac{1}{N}\sum_{i=1}^{N} VAR_{CO_2,i} \quad (3)$$

With the high-resolution (0.5ºx0.625º) 3-hourly GEOS-5 simulation results for 2018 from NASA Goddard Space Flight Center (Weir et al., 2021), we calculate the variance of atmospheric CO_2 concentration within each 2°×2.5° grid cell at every 3-hour interval. Then, we sample the CO_2 variance value ($VAR_{CO_2,i}$) at the grid cell containing the ith observation and the time closest to the observation. Subsequently, the monthly mean values of the *N* co-sampled variances are derived (ERR^2_{REP})."

20. Line 202. Please explain, "the regional average of error matrices."

[Response] we revised that sentence as follow:

"By applying 1000 bootstrap resampling to the monthly grid-based error statistics (e.g., *RMSE*, ERR_{MIP}, ERR_{REP}, and ERR_{TOT}) within each region, we obtain regional mean values of these error statistics, along with their corresponding 95% confidence intervals."

21. Figure 2a. Some regions have very, very few observations. What does that do to your results?

[Response] The availability of observation data in each region impacts both the reliability of our error statistics for quantifying regional flux errors and the area extents represented by these statistics. The monthly true flux error ($h(err_{f_t})$) is calculated using the Eq. 9, $h(err_{f_t})^2 - h(err_{f_e})^2 = RMSE^2 - ERR^2_{TOT}$. Out of 181 cases, representing the total months of observation across all seven regions, $h(err_{f_t})$ can be derived using this equation in 158 cases. However, in 23 cases (13% of total cases), $h(err_{f_t})$ cannot be derived from this calculation method when ERR_{TOT} and/or $h(err_{f_e})$ values fell outside the applicable range (Figure 5a-g). Around 40% of the exception cases occur in South America where observation data cover 1-6 grid cells by month. This indicates that observation data are insufficient to quantify the monthly flux errors in this region. In addition, the limited data availability results in a larger uncertainty range of the ratios of three-year mean $h(err_{f_e})$ to $h(err_{f_t})$, which are key metrics for quantifying regional terrestrial biosphere flux errors (Figure 5h). For example, the uncertainty ranges of the 95% confidence interval are 0.51, 0.53 and 1.05 for $h(err_{f_e})$ to $h(err_{f_t})$ ratios in Europe, Australia, and South America respectively, while the uncertainty ranges are 0.28 and 0.22 in mid-latitude north America and East Asia respectively, where observations cover wider areas and occur more frequently. Lastly, to identify areas that primarily contribute to the computed three-year mean error statistics, we considered regions that were selected as effective areas for at least eight months or more (Figure 6a; outlined area). Our error statistics computed from data in mid-latitude North America and East Asia represent broad regions encompassing the United States and eastern parts of East Asia. In contrast, those computed from data in Alaska and Europe, where observation made for limited periods and at specific locations, represent much smaller areas.

These results highlight the importance of frequent airborne measurements with extensive spatial coverage for reliable quantification of errors in regional terrestrial flux estimates derived from inverse models. We included above explanations in the revised manuscript (lines 275-279, 448-450, and 513-527).

22. Lines 312-313. Are the RMSE values between 1 and 3 ppm? Or is 1-3ppm the range of the values of RMSE?

[Response] We revised that sentence as follow:
"RMSE values in all these regions exhibit significant monthly variations, **with values falling within the range of 1-3 ppm,** with no clear seasonality possibly due to variations in observation routes (Figure 4)."

23. Figure 4. caption. I think these are monthly values of RMSE. Monthly variations of RMSE sounds to me like the variance of the RMSE.

[Response] we revised that expression in Figure 4 and Figure 5 as follow:

"Figure 4: (a-g) **Monthly values** of RMSE, …"
"Figure 5: (a-g) **Monthly values** of $h(err_{f_e})$ and …"

24. Paragraph starting on line 336. What is the main point of this paragraph? I have the same concern for all the paragraphs up to line 384. These paragraphs tend to describe the contents of the figures. It is hard for me to extract the main result. I suggest starting each of these paragraphs with a topic sentence that presents your main finding, then use the paragraph to explain this finding.

[Response] We appreciate the reviewers' comments. We revised these paragraphs shown in line 374–415 in the revised manuscript.

25. Line 386. I cannot find in section 2 where to find the method for determining the most influential areas for observed atmospheric CO2. And again, this is not a result.\

26. Line 393-394. I do not understand what is meant by the sentence starting with "Figure 6a…" and I don't understand the associated figure. Further, the text following this statement describes methodology, not results. Can you please explain Figure 6 methodology in the methods section of the text?

[Response] We agree with the reviewer's comments regarding the paragraph containing sentences (Line 386, Line 393-394), which described the methodology. We have relocated this content to the methodology section and revised the sentences to better convey our approach. The methods for determining the most influential areas for observed atmospheric CO2 and for deriving Figure 6 are now described in lines 297–322 in the revised manuscript.

27. Lines 419-421. I don't understand how this follows from the preceding text. If this is the main finding, please begin the paragraph with this statement, then use the paragraph to explain this statement. At present, I cannot follow this argument. It is an interesting argument. Please explain it more clearly.

[Response] We revised the paragraph by reorganizing the sentences and incorporating information on anthropogenic emissions as follows:

"Finally, by using the three-year regional mean ratios between $h(err_{f_e})$ and $h(err_{f_t})$, we compute the true errors in the annual terrestrial fluxes over the effective areas averaged for the period 2015–2017 (Figure 6). We find that the actual terrestrial biosphere flux errors are underestimated, particularly in regions where annual CO_2 emissions from fossil fuel combustion exceed annual terrestrial biosphere fluxes by 3-31 times. The airborne measurements carried out in mid-latitude North America, East Asia, and Southeast Asia are influenced by a broad region encompassing the United States, the eastern part of East Asia, and the western part of Southeast Asia where fossil fuel CO_2 emissions are 1,341, 2,443, and 815 Tg C year^{-1}, respectively. The first two regions are estimated as significant terrestrial biosphere CO_2 sinks, with estimated fluxes of −414 ± 279 (ensemble mean ± 1σ) and −561 ± 380 Tg C year^{-1}, in contrast to Southeast Asia (26 ± 118 Tg C year^{-1}). However, the CO_2 sinks are more than 3 and 4 times smaller than the fossil fuel CO_2 emissions, respectively. The recalculated terrestrial biosphere flux errors in these regions exceed the ensemble spread with values of 374, 643, and 211 Tg C year^{-1}. Observations in Europe and Australia, conducted over limited periods and specific locations, mainly represent certain areas in the western Europe and the southeastern part of Australia, where fossil fuel emissions (234 and 53 Tg C year^{-1}, respectively) are around four and five times greater than terrestrial biosphere sinks (−51 ± 34 and −10 ± 67 Tg C year^{-1}). The recalculated terrestrial biosphere flux errors in these regions are also larger than the ensemble spread, estimated at 65 and 114 Tg C year^{-1}, respectively. On the contrary, the most influential areas for the observation in Alaska and South America, encompassing the southeastern region of Alaska and the northern part of Brazil, characterized as a terrestrial biosphere sinks of −8 ± 11 Tg C year^{-1} and sources of 625 ± 387 Tg C year^{-1}, respectively, which are comparable to or more than 10 times greater than fossil fuel emissions (10 and 38 Tg C year^{-1}). The observation-based estimates of true terrestrial biosphere flux errors are almost identical to the ensemble spread in both regions with values of 11 and 398 Tg C year^{-1}, respectively."

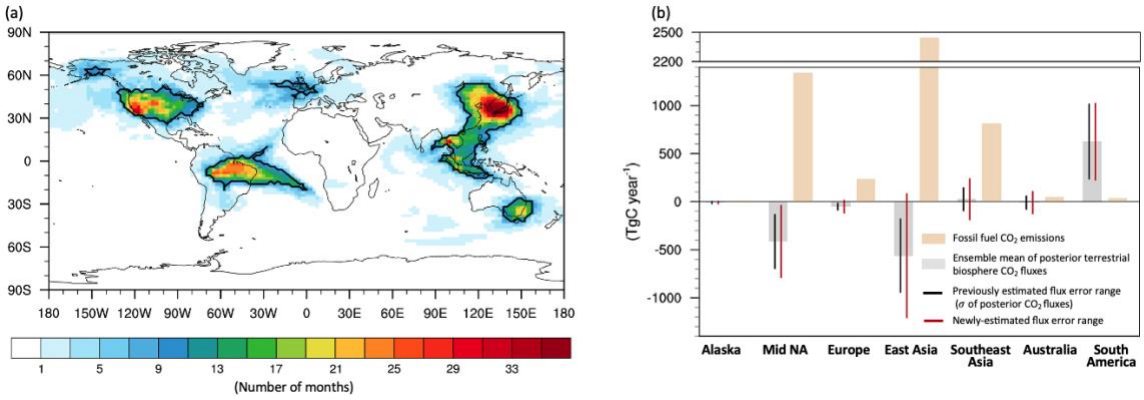

Figure 6: (a) Number of months selected as the effective area for airborne measurements. The outlined area represents selected areas for more than eight months or equal. (b) Annual total terrestrial biosphere CO_2 flux obtained from the ensemble mean of ten OCO-2 MIP models and annual total fossil fuel CO_2 emissions estimated from ODIAC data for each outlined area averaged over the period 2015–2017. The error bars in black and red indicate the one standard deviation of the inversion estimates and the newly estimated error range from this study, respectively.

28. Lines 424-428. This is material for the introduction, not the discussion.

[Response] We delete these sentences in the revised manuscript.

29. Line 435-437. This sentence needs work.

[Response] We revised that sentence as follow:
"For example, although the three-year mean errors in representation and transport in East Asia exceed those in Southeast Asia by 0.5 and 0.3 ppm, the disparity in projected mean true flux errors onto CO_2 space between the two regions is only 0.2 ppm."

30. Line 437. This result could be a natural consequence of what?

[Response] We revised that sentence as follow:

"This result is supported by previous studies highlighting that the spatial distributions of simulated CO_2 concentrations can vary significantly depending on the transport model (Schuch et al., 2023) and their spatial resolution (Stanevich et al., 2020)."

31. Line 444. I don't follow the "This underestimation…" sentence. Please clarify.

32. Line 444. What do you mean by "the common assumptions and observations…"? Are you arguing that since many ensemble members share common data and common methodological assumptions, this results in the spread among them being an underestimate of the true uncertainty in fluxes? This is possible and an interesting assertion, but I don't think it is proven by this work.

[Response] We agree with the reviewer' comment. This study does not specify major reasons for the underestimation in true flux errors, but discusses possible causes for the underestimation. We revised that sentence to clearly convey our intention as follow:

"The underestimation of true flux errors can arise from multiple factors, posing a challenge in determining main cause of the underestimation. Possible reasons include errors in methodological assumptions and atmospheric CO_2 observation data commonly applied to all OCO-2 MIP ensemble members because flux errors arising from these components are not captured by the ensemble spread."

33. Line 449. What is a "main source region"?

[Response] We revised it as follow:

"The underestimation of true flux errors only in regions **with more than three times greater fossil fuel emissions than biosphere fluxes** suggests …"

34. Lines 458-459. I don't understand the origins of the 15% figure, or the meaning of "challenges" in estimating monthly flux errors. I very much agree with the concern at the end of this paragraph that areas with limited data

may not have sufficient data for computing reliable error statistics for the flux inversions. I think these are related topics. Please clarify.

[Response] Right. This paragraph discusses the impact of regionally different observation data availability on our results. We revise that paragraph as follow:

"The reliability of our observation-based regional flux error estimates is based upon the data availability of airborne measurements. Although our approach is generally effective in estimating a regional mean of monthly $h(err_{f_t})$, it is not applicable in 15% of our total cases (shown in Figure 5), when measurements were mostly made in local areas covering one to six 1°×1° grid cells within each region. This limitation may be attributed to the application of a common method for calculating observation errors across all data points, which might not adequately identify specific outliers. Caution is required when applying our approach to monthly-scale analysis, especially when using observations made locally. Extending the calculation period to several months or longer (e.g., Figure 5h) is a suitable strategy for mitigating the impact of outliers and obtaining more robust results. In fact, the ratios of three-year mean $h(err_{f_e})$ to $h(err_{f_t})$, which are key metrics for quantifying regional flux errors (Figure 5h), have a smaller uncertainty in mid-latitude North America and East Asia where wide and consistent airborne data are available, than over Europe and South America, where aircraft observations are sparse and only have intermittent data coverage. In addition, it is noteworthy that the $h(err_{f_e})$ to $h(err_{f_t})$ ratios derived from continuous observations enable the computation of unbiased true errors in the average annual ensemble terrestrial fluxes for the analysis period, compared to those from limited observation periods (e.g., in Alaska). These results highlight the importance to have frequent airborne measurements with extensive spatial coverage for the reliable error quantification of regional terrestrial flux estimates derived from inverse models."

35. Line 477. "Second…" This is another paragraph.

[Response] We separated the paragraph in the revised manuscript.

36. Line 491-492. I am not convinced that these flux errors are largest in errors with large anthropogenic fluxes. This is a plausible hypothesis, but I would not say that the results reveal this to be true. I would like to see a more careful analysis of the fossil fluxes in the relevant influence regions, and the relationship between the strength of fossil fluxes and these flux errors to be convinced.

[Response] We agree with the reviewer's comment. We add information on fossil fuel emissions to support our main finding, "the actual errors in ensemble mean of annual terrestrial biosphere flux estimates of OCO-2 MIP are underestimated, particularly in regions with higher fossil fuel CO2 emissions compared to terrestrial biosphere CO2 fluxes" in the result part in the revised manuscript (lines 452–471; or please refer to our response to your 27th comment).