Reply to Reviewer 1:

General comments

This paper presents a new algorithm for ingesting information about surface CO2 fluxes from bottom-up estimations into a top-down atmospheric inversion system. The main innovation is that instead of using the bottom-up fluxes as a priori fluxes in the inversion, the spatial gradients of the fluxes are instead assimilated as observations, which the authors argue increases the signal-to-noise ratio and avoids contaminating the analysis with biases in the bottom-up fluxes.

While the idea of assimilating the spatial flux gradients rather than the fluxes themselves is interesting and worth investigating, I had a hard time assessing the manuscript on scientific quality and validity because of incomplete information, confusing use of terms, and distracting issues in the text and figures. More detailed comments are provided below. I believe addressing these issues goes beyond the scope of a major revision and recommend the authors to substantially rework the manuscript in terms of content and presentation. Only then can a proper assessment be made about the scientific quality.

Reply: Many thanks for your constructive comments and suggestions. We acknowledge that we have missed some important information and the experiments are not optimal. We are sorry for the inconvenience. We have reworked the manuscript in terms of method, experiment setup, and results. And the manuscript was also polished by AJE. The main changes are as follows:

1. Information: Adding more details about the short assimilation window and long observation window / experiment setup of initial condition, ensemble size etc. / observation localization / additive inflation / generation of pseudo observation.

2. Clearer method description: a) Assimilating CO_2 observation before assimilating a priori (Fig. 1). b) Using $\nabla \mathbf{f}$ to represent the spatial gradient.

3. New experiment setup: a) OCO2+insitu vs insitu only. b) Making sure the weight of a priori is identical in different experiments. c) Changing CASA to SiB4 as the a priori.

4. Clearer message from the OSSE results: Better hemispheric flux estimates using $\nabla \mathbf{f}$ in both experiments of assimilate OCO2+insitu or insitu only.

We hope that your major concerns are clarified and addressed. Note that we have made large revision on this manuscript that some parts of this paper are deleted or replaced (e.g., results section). Some sentences you commented on may be deleted. We are sorry about that, but these modifications are mainly based on your major concerns.

1. Incomplete information

Many important pieces of information are missing from the manuscript, which makes it difficult to follow and fully understand the experiments and results. Sometimes the reader is referred to previous papers (e.g., Liu et al. (2019, 2022) and Kang et al. (2012)), and at other times it is not clear where to find the missing information. I am not saying that the authors need to explain everything in this manuscript— for example, I think it is perfectly fine to refer to other papers for details about the LETKF algorithm— but there are many cases where the lack of information hinders the reader's comprehension of the present study. Often this information can be provided in one or two short sentences, so text length should not be a problem. Here are just some examples of information that I miss in the current manuscript:

1. Ensemble size, and if the results are sensitive to ensemble size.

Reply: Thanks for the comment. We have added the details. The ensemble size used in COLA is 20 (Line 96). It has been tested in Liu et al. (2019). And this setup is also tested in real data assimilation experiments (e.g., the OCO2MIP-v10).

2. What is the background (both fluxes and concentrations) at t=0 is. How the ensemble members were perturbed at t=0 (i.e., what is the assumed initial background error covariances?).

Reply: Thanks for the comment. We have added the details. The ensemble initial conditions for the assimilation run are randomly selected from the control run from 15 August to 15 September 2014 (Line 190). This is similar to the additive inflation process used in COLA.

3. The time resolution of the assimilated observations. The text says "For satellite data, we used a 10-second averaged …", but I assume that the system did not actually assimilate 10-second satellite observations? Were the observations averaged to daily values before being assimilated? For in situ observations, were observations from all hours used, or only e.g. measurements that were identified as being representative of large well-mixed air masses?

Reply: Thanks for the comment. We have added the details of pseudo observations to the appendix section (Text A2, Line 358). We discussed and decided not to put the details in the main text because it is not the key points of this paper and may distract readers.

The time resolution of the postprocessed real OCO-2 observation is 10-second. Since the timestep of GEOS-Chem is configured to 30 minutes, we can not sample from the nature run at 10-second timestep. Thus, the pseudo OCO-2 observation is sampled at 1-hourly timestep (Line 371) but using the real location of OCO-2 at 10-second timestep.

The insitu observations are averaged to 1-hourly at each station, and only the data from 14:00 to 16:00 local solar time are used (Line 360). The insitu network are selected from the GV obspack that are not affected by local sources (Fig. A1a).

4. The manuscript should be explicit about OCO-2 observations being column-integrated CO2 (XCO2) and explain how the synthetic XCO2 observations were calculated from the truth simulation.

Reply: Thanks for the suggestion. We have added the details of the generation of XCO_2 observations (Line 374, Equation 13).

5. The text refers to Liu et al. (2022) for information about the observation errors, but it would be nice if a short description could be provided here as well.

Reply: Thanks for the suggestion. The measurement error is obtained from the ObsPack. And the representative error is created based on half the variance of the deseasonalized and detrended CO_2 time series at a given station (Line 363, Chevallier, 2010).

Reference: Chevallier et al., 2010, CO2 surface fluxes at grid point scale estimated from a global 21 year reanalysis of atmospheric measurements.

6. The inflation method is not sufficiently explained. I assume the mean of the additive perturbations were subtracted to not influence the ensemble mean (similar to what was done in Liu et al. (2022))? Were the magnitudes of the perturbations also scaled to a predefined value, and if so, what is the value?

Reply: Thanks for the suggestion. We have added the details of inflation method to the appendix section (Text A1, Line 313). Also, we discussed and decided not to put the details in the main text because it is not the key points of this paper and may distract readers. Yes, the mean of the additive perturbations was subtracted.

Similar to Liu et al. (2022), we maintained the additive inflation strategy but from an adaptive perspective. The ensemble uncertainty reduction differs in different regions with different spatial coverage of CO_2 observations. Typically, the reduction is expected to be small over poorly observed tropics. Thus, adding randomly selected fields to the ensembles directly may override the ensemble

characteristic and overinflate the ensemble spread in those areas. Thus, we predefined a time varying ensemble spread scale. This scale is calculated based on the temporal variability of the a priori F_{TA} (SiB4). The added anomaly fields are scaled before added to ensure that the inflated ensemble spread does not greater than the predefined scale. More details can be found in the Text A1.

7. The text mentions a short (1 day) and a long (7 days) assimilation window. How were the parameters adjusted within the long assimilation window? Liu et al. (2022) states that the parameters were adjusted on a daily basis within the long assimilation window— if it is the same in this study, it should be explicitly stated. Or are the parameters assumed to be constant within the assimilation window? What are the temporal error correlations for the parameters within the same assimilation window, and do they allow the parameters to be adjusted differently on a daily basis?

Reply: Thanks for the suggestion. We have added more details on the window setup in the Method section. And we explictly guide readers to read more about the setup in Liu et al. (2019). The parameters are assumed to be constant within the long observation window. Thus, the parameters of different days within the observation window are identical.

The text decribe the windows are (Line 94): *"COLA implements a new version of LETKF with an advanced feature of a short assimilation window (1 day) and a long observation window (7 days) to enhance the SCF estimation (Liu et al., 2019). In this approach, the persistent flux parameter and dynamic CO_2 state are updated on a daily basis using observations within the long observation window (7 days). To achieve that, for each analysis cycle, GEOS-Chem forecasts for 7 days to generate forecast observations, which are then assimilated with the corresponding observations using LETKF. This optimization process updates the model state (CO2) and parameters (SCFs) at the end of the assimilation window, serving as the forecast initial conditions and timing for the subsequent analysis cycle. A comprehensive description of this unique LETKF feature can be found in the work of (Liu et al., 2019)."*

8. What is the temporal resolution of the flux products? From my understanding VEGAS can simulate daily fluxes, but what about CASA, are those monthly fluxes? If so, how were they downscaled to daily values?

Reply: Thanks for the comment. We are sorry that we missed the information. We have added the details. VEGAS can simulate daily fluxes. Before you mentioned the temporal resolution of CASA, I did not know that CASA can only simulate monthly fluxes, and it is downscaled to daily values based on meteorology data. In the last version we downloaded the preprocessed daily CASA data from GEOS-Chem official website. Considering that CASA is very different from VEGAS and is not a DGVM but a satellite driven model, we replaced the CASA with SiB4 as the a priori F_{TA} used in this study. And SiB4 can simulate the daily fluxes.

The revised text contains the temporal information of a priori F_{TA} (Line 185):" *Another terrestrial model of SiB4 (Haynes et al., 2019) is used to provide an independent daily bottom-up F_{TA} estimation for the a priori regularization in the assimilation experiments*".

9. Were the fluxes/spatial gradients of fluxes assimilated before or after the CO2 observations? Does the order matter?

Reply: Thanks for the comment. The fluxes/spatial gradients of fluxes are assimilated after the CO_2 observations. We have revised the text and Figure 1 to show the order. Since both are treated as observations, we separated into two steps in order to assess the impact of the a priori on ensemble

uncertainty reduction (Fig. R1). We have tested to assimilate the a priori simultaneously or after, the order does not influence the results.
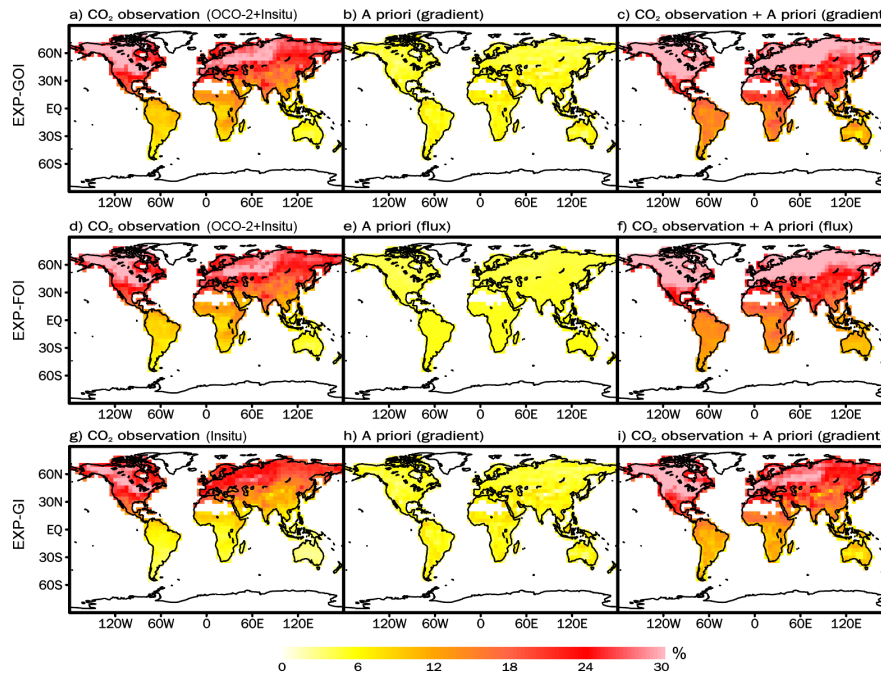

Figure R1: a-c) The ensemble uncertainty reduction in EXP-GOI from the satellite and in-situ CO₂ observations, the a priori of the gradient, and the sum of the former two. d-f) The ensemble uncertainty reduction in EXP-FOI from the satellite and in-situ CO₂ observations, the a priori of the original flux, and the sum of the former two. g-i) The ensemble uncertainty reduction in EXP-GOI from the in-situ CO₂ observations, the a priori of the gradient, and the sum of the former two.

This is not an exhaustive list, and I do not think it should be the reviewers' job to list everything that should be covered. The authors should think carefully about what information is relevant to include in the manuscript.

Reply: Thanks for your careful reading and we are sorry for the inconvenience. We have added other necessary information to support the key points.

References

Kang, J.-S., E. Kalnay, T. Miyoshi, J. Liu, and I. Fung (2012): Estimation of surface carbon fluxes with an advanced data assimilation methodology. Geophys. Res. Atmospheres, 117, D24101, doi:10.1029/2012JD018259.

Liu, Y., E. Kalnay, N. Zeng, G. Asrar, Z. Chen, and B. Jia (2019): Estimating surface carbon fluxes based on a local ensemble transform Kalman filter with a short assimilation window and a long observation window: an observing system simulation experiment test in GEOS-Chem 10.1. Geosci. Model Dev., 12, 2899–2914, doi:10.5194/gmd-12-2899-2019.

Liu, Z., N. Zeng, Y. Liu, E. Kalnay, G. Asrar, B. Wu, Q. Cai, D. Liu, and P. Han (2022): Improving the joint estimation of CO2 and surface carbon fluxes using a constrained ensemble Kalman filter in COLA (v1.0). Geosci Model Dev, 18, doi:10.5194/gmd-15-5511-2022.

2. Confusing use of terms

There are several terms that are used in a way that is confusing or add little to no value. Again, here are just some examples:

1. "Unique" is used throughout the manuscript, for example, "as a unique observation" (Line 21), "a unique feature of a short assimilation window" (Line 80), and "Another unique feature" (Line 82). I do not see the point of stating that these features are unique. For assimilating a priori information as observations, this has been done before (a recent example is Kaminski et al., 2022). For the first case, maybe the authors meant "as a special observation" or something similar? Similar to the first point, "comprehensive" is often used (Lines 22, 59, 111, 292), but it is not clear in what sense the described objects are comprehensive.

Reply: Thanks for the comment. We have revised the corresponding words through the manuscripts. The "unique observation" is revised to "special observation" through the manuscript (e.g., Title). The "a unique feature of a short assimilation window" is revised to "an advanced feature of a short assimilation window" (Line 89). "Another unique feature" is revised to "Another notable feature".

Thanks for pointing out that there is other similar approach. We have presented the AAPO method at the OCO2MIP telecon and Dr. Andy Jacobson also mentioned that the TransCom studies have been added the a priori flux into the observation vector. Thus, it is indeed not a unique method. And using "special" is more appropriate.

We use "comprehensive" in the manuscript to emphasis that the AAPO method also using the similar data assimilation approach as assimilating the CO_2 observations. Thus, both a priori and CO_2 observations are assimilated. We think using "comprehensive" may mislead the readers or add no value, and we deleted the words correspondingly.

2. I found it confusing that fluxes from the CASA model are described as a priori fluxes even though they were not used as a prior in the inversions. Maybe refer to them as "bottom-up fluxes" instead?

Reply: Thanks for the comment. The CASA (now SiB4) modeled fluxes is used as a priori in some experiments (EXP-FOI and EXP-FI). And the two experiments are compared with the other two experiments (EXP-GOI and EXP-GI) that used the spatial gradient of SiB4 in order to show the advantages of spatial gradient.

3. Many acronyms are confusing. For example, in Summary and Discussion, AAPO is redefined on Line 290, but it is not clear in this context how the acronym was formed.

Many equations introduce new acronyms, for example TSG and PSG in Equation (5), and ASG in Equation (8). How about using a different variable (e.g., ∇f) or a subscript (e.g., "SG") for the spatial gradient of the fluxes, and then superscripts A, T, and P for analysis, truth, and prior, respectively? This would reduce the number of acronyms the reader has to keep track of.

Equation (6) introduces the "S" superscript, which I assume stands for "SCF", but it is not defined or used anywhere else in the manuscript, which is confusing. I would suggest dropping the superscript, or have "SG" and "f" as subscripts in Equations (5) and (6) instead. Generally, try to be consistent throughout the manuscript.

Similarly, what do "NP", "P", "ASG", and "AP" in the experiment names stand for? (I have my guesses, but it should not be the reader's job to figure this out.)

Reply: Many thanks for the constructive suggestion. The acronyms are misleading and not informative. We have deleted some unnecessary equations and carefully revised the acronyms in the equations. We have checked the acronyms that are redefined or not clearly defined.

The spatial gradient of flux is represented by $\nabla \mathbf{f}$, while the flux itself is represented by $\mathbf{f}$ (Line 205, Equation 3, 4). The truth, analysis, and a priori are represented by T, a, and Ap, respectively. The meaning of experiment names is changed to represent the a priori and CO_2 observations and described in section 3.1. For example, EXP-GOI stands for assimilates the Gradient and the synthetic OCO-2 and in-situ observation (Line 191, Table 1).

4. Line 138 says that "COLA is a flow-dependent ensemble-based DA". The term "flow-dependent" makes sense in the context of e.g. atmospheric data assimilation, where the errors depend (and often follow) the atmospheric flow, but it is not clear what exactly it means here. Given that there is no dynamic model for the parameters, how is the system flow-dependent?

Reply: Thanks for the comment. We use "flow-dependent" to emphasis that COLA fully uses the persistence from the last assimilation time as the first guess of the current assimilation time. This may mislead readers and is not appropriate. And I agree that flow-dependent make sense with dynamic model. Thus, we deleted the corresponding expression.

5. The text uses "significant/significantly" in many places. I would recommend reserving these words for describing statistical significance when comparing results. If the authors performed statistical significance testing, they should state what tests were used and the significance level.

Reply: Thanks for the suggestions. We have made large revision on the results. Following your suggestions, we avoid using "significant/significantly" when comparing two experiments but using ratio/bias to show the difference and improvement. For example, in Line 280, "*In the North American Temperate and Europe, where the observations are dense, ..., and the RMSEs in EXP-FOI are larger than those in EXP-OI by 59.2% to 91.5%*".

6. Lines 289–290 state that "the spatial gradient of a bottom-up model estimation is dynamically assimilated". What is "dynamically" referring to here?

Rely: Thanks for the comment. In this context, we use "dynamically" to emphasis that the spatial gradient calculation can partly compensate the deficiency of lacking a dynamic model. This is not an accurate expression. Thus, we deleted "dynamic/dynamically" in the title and main text corresponding.

References

Kaminski, T., M. Scholze, P. Rayner, M. Voßbeck, M. Buchwitz, M. Reuter, W. Knorr, H. Chen, A. Agusti-Panareda, A. Löscher, and Y. Meijer (2022): Assimilation of atmospheric CO2 observations from space can support national CO2 emission inventories. Environmental Research Letters, 17, 014015, doi:10.1088/1748-9326/ac3cea.

3. Language and figures

The language could be improved in several instances, for example:

1. Articles are sometimes missing or incorrectly added, or the noun should be plural instead of singular or vice versa. Here are just a few examples: "and a priori flux from …" -> "and a priori fluxes from" (Line 15), "weigh a priori flux to the background" -> "weigh a priori fluxes to the background" (Line 16), "because of the deficiencies in understanding" -> "because of deficiencies in understanding" (Line 19), "LETKF estimates SCF as …" -> "LETKF estimates SCFs as …" (Line 76).

Rely: Thanks for finding the mistakes. We have revised the corresponding words you mentioned and other mistakes throughout the manuscript. And it is polished by AJE.

2. The text often uses subjective words such as "good" and "better". I would suggest using more objective and descriptive words. For example, instead of saying "produces much better

estimation …”, I would say something like “produces more accurate flux estimates in terms of x and y”.

Reply: Thanks for the suggestions. In the revised results section, the comparison between experiments are not described by descriptive words. For example, in Line 261, “*However, the global RMSE in EXP-GOI is 26.4% smaller than that in EXP-OI, indicating that the gradient is better than the original flux in estimating the global seasonal cycle*”.

3. When describing what was done, I suggest using past tense (“GEOS-Chem was run” (Line 160), “Four experiments were performed” (Line 205), etc.) rather than present tense.

Reply: Thanks for the suggestions. We have checked and revised the tenses.

The figures are not the clearest. Here are a few suggestions for improvements:

1. Figure 2: The blue-red colorbar has similar lightness at the upper/lower ranges of the colorbar, especially for the red part, which makes it hard to see the difference between e.g. 2.0 and 1.6. Maybe use a discrete colorbar, and change to a colormap with a larger range of lightness.

Reply: Thanks for the suggestions. We have revised the colormaps in map figures with more and clearer levels. In Figure 2, the colormap is changed from blue/white/red to blue/skyblue/white/orange/red.

2. Figure 3: It is sometimes hard to discern the “truth” line in the top panel. Maybe make the black line thicker.Also Figure 3: It would be easier to read the plot if the RMSE/BIAS values were listed in the same order as the legend showing the experiments.

Reply: Thanks for the suggestions. In Figure 3 (now Figure 4), we have made the “truth” line thicker. The RMSEs are presented following the order of legend.

4. Concerns about the experimental setup

I have several concerns about the experimental setup, which could affect the interpretation of the results. My main concern is that the assumptions are too idealistic to be applicable in real data cases. In particular, the OCO-2 observations in these perfect model OSSEs will directly reflect the surface CO2 fluxes and therefore provide a strong constraint on the spatiotemporal flux variations. In the real world, XCO2 observations are generally sensitive to other error sources such as systematic errors, representativeness errors (the swath width of OCO-2 is much smaller than the model grid points here and the satellite measurements are also affected by e.g. clouds), errors in the atmospheric CO2 background, and atmospheric transport errors. As a first step, how were representativeness errors included in the synthetic observations? Did the authors assume that there was an XCO2 observation in a 4°×5° model grid cell if there were more than x% XCO2 measurements in the grid cell? If so, what was the x threshold?

Reply: Thanks for the comment. We understand your concern about the observations. Since this study is aimed at validating the AAPO method in the context of OSSE, the effect of AAPO may be mixed with other factors if we consider other sources of uncertainty/error. And the systematic error of satellite retrievals is not fully understood currently (Taylor et al., 2023), it is hard for us to add such error to the pseudo observations. However, the generation of pseudo OCO-2 observations do add a random error to it (Text A2). The pseudo OCO-2 retrieval is created based on the following equation,

$$XCO_2^{pseudo} = \mathbf{p} \cdot ((\mathbf{I} - \mathbf{ak}) \cdot \mathbf{c}^{prior} + \mathbf{ak} \cdot \mathbf{c}^{nature}) + \varepsilon,$$

where XCO_2^{pseudo} represents a pseudo OCO-2 retrieval; $\mathbf{p}$ is the vertical pressure weight profile corresponding to the averaging kernel profile $\mathbf{ak}$; $\mathbf{c}^{nature}$ is the a priori CO_2 vertical profile prescribed by the ACOS retrieval algorithm; $\mathbf{c}^{nature}$ is the simulated CO_2 vertical profile sampled from the nature run; $\mathbf{I}$ is an identical vector; and ε is the added zero-mean random uncertainty based on the observation error. Baker et al. (2022) described how the error is estimated during the 10-second averaging process.

Since the OCO-2 data provided by Baker et al. (2022) is a 10-second averaged product, and it has been exstensively used in the OCO2MIP inversions. The distance between two adjacent observation is ~67.5 km (~0.6 degree). Thus, there are approximately 6 observations within a model grid cell, and there are around 80 observations in some areas annually (Fig. A1b). Following the "assimilate_flag" described in postprocessed OCO-2 data, only good quality observations in the real data are used to generate the pseudo observation in the OSSE.

We understand your concerns, since the spatial resolution of OCO-2 is very small compared with model resolution, it is not practical to fully reproduce the characteristics of the real observation. But the added random errors are independent from each other. And, similar approach of generating pseudo satellite observation has been applied in published researches to exam their inversion systems (e.g., Liu et al., 2014;)

References:

TE Taylor et al., 2023, AMTD, Evaluating the consistency between OCO-2 and OCO-3 XCO estimates derived from the NASA ACOS version 10 retrieval algorithm.

DF Baker et al., 2022, GMD, A new exponentially decaying error correlation model for assimilating OCO-2 column-average CO_2 data using a length scale computed from airborne lidar measurements.

To make it easier to understand the results and compare them with other established inversion systems, I suggest the authors perform another set of OSSEs where only in situ observations are assimilated.

Reply: Thanks for the suggestion. We have made two extra experiments (EXP-FI and EXP-GI) that only assimilate the insitu observations. Comparing between the two experiments, we find that the performance of EXP-GI that using the spatial gradient is also better than that in EXP-FI.

I am also wondering if some of the results could be due to the specific algorithms in COLA rather than indicative of the performance of general inversion systems. For example, the uncertainty of the assimilated bottom-up fluxes/flux gradients were assumed to be proportional to the analysis uncertainty. Thus, if I understand this correctly, the data assimilation system will place more weight on the assimilated bottom-up fluxes/flux gradients compared with the CO2/XCO2 observations the more confident it is in its analysis. Is this a reasonable assumption? It seems to me that this would bias the EXP-P experiment toward the bottom-up fluxes, which seems to be what happens if you look at Figure 4, given that the analysis uncertainty will likely be reduced a lot when assimilating direct observations of the fluxes.

Reply: Thanks for the comment. Since there are some configurations/method are only applied in COLA, we acknowledge that the results may be affected by them. For example, the window length of COLA is relatively shorter than some long window-based inversion systems. We indicated that COLA may not be bettern than those system in in-situ only inversion (Line 247).

However, the uncertainty of a priori is assumed to be proportional to but 5 times larger than the analysis uncertainty, which means that the weight of a priori is assumed to be small. And the a priori only serves to regularize the ill-posed problem and does not overwrite the information given by the CO_2 observations. The weight of a priori and CO_2 observation can be expressed by the ensemble uncertainty reduction (EUR), which is defined as follows,

$$\mathbf{EUR}_{i,t} = \frac{\sigma_{i,t}^b - \sigma_{i,t}^a}{\sigma_{i,t}^b},$$

where $\boldsymbol{\sigma}_{i,t}^b$ and $\boldsymbol{\sigma}_{i,t}^a$ is the first guess and final analysis ensemble uncertainty, respectively, at a given grid point i and a given time t. Since there are two types of observation, the EUR can be separated to two parts of EUR^{CO2} and EUR^{Ap} as,

$$EUR_{i,t}^{CO2} = \frac{\sigma_{i,t}^{b} - \sigma_{i,t}^{a*}}{\sigma_{i,t}^{b}},$$

$$EUR_{i,t}^{Ap} = \frac{\sigma_{i,t}^{a*} - \sigma_{i,t}^{a}}{\sigma_{i,t}^{b}},$$

where the superscript $a*$ denote the analysis after assimilating CO_2 observation; the superscript Ap denotes the a priori.

As show in Figure R1, for EXP-GOI, the EUR^{CO2} in the northern middle and high latitudes of North America and Europe, where the CO_2 observation network is dense, can exceed 30%. Generally, the EUR^{CO2} decreases from north to south. In South America and Africa, the EUR^{CO2} is around 10%. Since the uncertainty of a priori is set to be proportional to the ensemble uncertainty, *EUR^{Ap} is approximately 5% and almost identical at different grids. For EXP-FOI, the EUR^{Ap} is configured to be identical to EXP-GOI; thus, the weights of a priori in EXP-GOI and EXP-FOI are expected to be identical.* However, in EXP-GOI, the EUR^{Ap} in the Qinghai-Tibet Plateau, Andes, Sahel, and Arabian Plateau is relatively larger than in the other areas, suggesting that, even though the error of the gradient at different grid points is identical, the EUR^{Ap} is also affected by the geography. In EXP-GI, which assimilated only the in-situ data, the EUR^{CO2} is approximately half that in EXP-GOI, resulting in similar EUR^{CO2} and EUR^{Ap} in South America and Africa.

For the additive inflation method, given that the ensemble flux deviations and the additive flux perturbations can both be positive and negative, is it possible that the ensemble spread will actually decrease when adding the perturbations?

Reply: Thanks for the comments on the additive inflation method. We have added a section to describe the inflation method (Text A1). Yes, since the flux perturbations are randomly selected from the a priori fluxes, it is possible that the ensemble spread will decrease when adding the perturbations.

Additionally, it is also possible that the ensemble spread will be overinflated. Because the EUR^{CO2} differs in different regions with different spatial coverage of CO_2 observations. Typically, the EUR^{CO2} is expected to be small over poorly observed tropics (Fig. R1, first column). Thus, adding randomly selected fields to the ensembles directly may override the ensemble characteristic and overinflate the ensemble spread in those areas. Thus, to partly overcome the decreased or overinflated problem, a time-varying ensemble uncertainty scale based on the temporal variability of the bottom-up estimation is defined before adding those selected fields. And the perturbations are rescaled based on the scale.

Another point about the inflation method: the daily fluxes vary mostly due to varying weather and advancing seasonal cycles (this is especially true if the fluxes were downscaled from e.g. monthly values). Thus, there are likely strong temporal correlations in the daily flux variations. I wonder if way of choosing additive inflation anomalies is optimal considering for example the subspace spanned by the ensemble members after adding the inflation perturbations. A quick check could be to perform a principal component analysis on the additive ensemble perturbations from the inflation method and see how the explained variance declines.

Reply: Thanks for the comments. We understand your concerns on the additive method. Since the paper is focused on the AAPO method, adding the details may distract readers from the main point and we put the details on the appendix section (Text A1).

We agree that there are strong temporal correlations in the daily flux variation, and the daily variations are what COLA looking for. Since The first guess of the SCFs is derived based on the persistence of the analysis from the previous assimilation cycle. The error of the first guess is the flux difference

of the current and the previous assimilation time. For example, the seasonal cycle is coherently evolving in the Northern Hemisphere and the added perturbation based on the additive method could partly express the spatiotemporal correlation. Since the inflation method approximates the covariance, it does not fully reproduce the real covariance structure.

Related to the previous point, would this scheme not put less emphasis on fluxes in regions with smaller seasonal cycles and/or less variations in weather (in particular temperature and solar radiation)? Is this desirable? It would be informative if there was a figure (for example in an appendix) showing some examples of the additive inflation flux anomalies, or the leading modes and their explained variance from the principal component analysis.

Reply: Thanks for the comments. Based on the additive inflation scheme, the ensemble spread in the tropics is smaller than that in the northern extratropics. However, this is not meaning that the scheme put less emphasis on the fluxes in the tropics.

If give a large ensemble spread in the tropics, the variation of the analysis fluxes could be larger than the temporal variation of the a priori fluxes and may lead to large temporal dipole noises. In some condition, the temporal variation of the a priori flux may be smaller than the true flux. To overcome this situation, we added an annual-mean ensemble uncertainty scale to the inflation process to ensure that the ensemble spread is not underestimated (Text A1, Line 308).

5. Comparison with traditional methods

The manuscript claims to compare the results from the new algorithm with a "traditional" method. Here the "traditional" method is basically Equation (2), whereby fluxes are forecast forward in time based on the analysis from the previous two time steps and the bottom-up fluxes at the forecast time. However, I would not call this the traditional method as (i) to my knowledge, this method is only used by CarbonTracker/CarbonTracker Europe and maybe one or two other systems; and (ii) the implementation here is not the same as the one in CarbonTracker or in the cited reference. For one thing, the dynamic model in CarbonTracker is applied on flux parameters, which are scaling factors on bottom-up fluxes, and not on the fluxes themselves (as in this manuscript). The idea is that the bottom-up fluxes capture the high-frequency variations in the fluxes, for example due to shifting weather, while the flux parameters correct for long-term systematic errors. It may therefore be reasonable to smooth the flux parameters in time and relax them back to the prior values. In this manuscript, however, Equation (2) would partly relax the fluxes back to the previous days' values, which would likely degrade the forecast fluxes in most cases (considering the advancement of the seasonal cycle and movement of weather systems) compared with the persistence model.

Reply: Thanks for pointing out the mistakes on describing the methods applied in CarbonTracker and CarbonTracker-EU. We have revised the corresponding sentences (Line 114): "*Consequently, in CO_2 inversion systems with longer assimilation windows, the initial estimate of carbon fluxes heavily relies on a priori regularization techniques. For example, Feng et al. (2009) utilized the a priori information as the initial guess for their EnKF calculations, while Peters et al. (2007) employed the a priori to derive carbon flux variation modes and corresponding scaling factors, which are optimized through EnKF*".

In the scheme of COLA (Line 109), the first guess of the SCFs is derived based on the persistence of the analysis from the previous assimilation cycle, which is considered a reasonable approximation for the current SCF estimate. This assumption of persistence aligns well with COLA's short assimilation window of 1 day. However, for systems with longer assimilation windows, such as those spanning several weeks for SCF estimation, the temporal dynamics of the terrestrial ecosystem,

including diurnal cycles and the transition from winter respiration to rapid spring growth, can introduce temporal lag and inaccuracies. Consequently, in CO_2 inversion systems with longer assimilation windows, the initial estimate of carbon fluxes relies more on a priori fluxes.

In conclusion, the persistence scheme applied in COLA relies on the short assimilation window of 1 day. As the reviewer pointed out that, if the window is long, the persistence is definitely not a good assumption and may lead to temporal lag.

This together with the concerns raised in the previous points makes it unsuitable in my opinion to refer to EXP-P as using "traditional" methods. I do agree that there should be a comparison with established CO2 inversion methods to assess the added value of the new algorithms introduced in this paper. Thus, I recommend the authors to set up a more "standard" inversion experiment and compare the results from this inversion with their results.

Reply: Thanks for the comments. We agree that EXP-P is not really a "Traditional" method. Since COLA is different from other traditional systems in the assimilation schemes, it is hard for us to make an experiment using real 'traditional' methods.

Thus, to make the key point clear, we decide to delete the experiment that directly add a priori fluxes to the first guess because it is not comparable to the experiment using the AAPO scheme in terms of the weight of a priori. And the experiments using AAPO methods modified each ensemble member and reduce uncertainty (Fig. R1), while EXP-P only modified the ensemble mean. Finally, the key point of this paper is focusing on the comparison between the flux and its spatial gradient.

Specific comments

Lines 18–19: "systematically biased at different spatiotemporal scales because of the deficiencies in understanding of some underlying processes". It could also be due to uncertain parameter values and/or initial conditions in variables related to the carbon cycle.

Reply: Thanks for the comments. We have revised to, "*However, the "bottom-up" flux estimations, especially the simulated terrestrial-atmosphere CO_2 exchange, are usually systematically biased at different spatiotemporal scales due to the limitation in understanding and parameterizing some underlying processes.*".

Lines 26–28: "We suggest that the AAPO algorithm can be applied to other greenhouse gas (e.g., CH4, NO2) and pollutant data assimilation studies." This was not shown in the study, so I suggest removing it from the abstract.

Reply: Thanks for the comments. We have deleted this sentence in the abstract.

Lines 35–36: "of the Bayesian synthesis (…) and data assimilation (DA) techniques". I wonder why the authors separate between Bayesian synthesis and data assimilation techniques here. Most data assimilation methods in widespread use are based on Bayesian inference (Rayner et al., 2019).

References

Rayner, P. J., A. M. Michalak, and F. Chevallier (2019): Fundamentals of data assimilation applied to biogeochemistry. Atmos. Chem. Phys., 19, 13911–13932, doi:10.5194/acp-19-13911-2019.

Reply: Thanks for the comments. We agree that this may lead misunderstandings. The sentence is revised to, "*The SCFs can be inferred from atmospheric CO_2 measurements using "top-down" (hereafter quotation marks will be omitted) techniques of the Bayesian synthesis (e.g., Rodenbeck et al., 2003; Zammit-Mangion et al., 2022; Cho et al., 2022) and, ensemble Kalman filter (e.g., Peters et al., 2007; Feng et al., 2009; Jiang et al., 2022; Liu et al., 2022), and 4-dimentional variational (4D-var) techniques (Chevallier et al., 2010; Liu et al., 2014).*"

Lines 37–28: "However, the top-down estimation could be ill-posed because of … and systematic errors of the transport model and satellite retrieval". I would argue that the most common reason for ill-posed problems in inverse modeling is that there is not a unique solution, which is mainly caused by a lack of observational constraint and insufficient regularization. I would not consider systematic errors to be the primarily cause of ill-posed problems. From a quick scan I could not find the term "ill-posed" mentioned in any of the cited references.

Reply: Thanks for the comments and suggestion. We agree that the systematic errors are not the reason of being "ill-posed" but other influence factors. The sentences are revised to be clearer in Line 41, "*However, the top-down estimation could be ill-posed because of the sparseness feature of atmospheric CO_2 observations. And systematic errors in the transport model and satellite retrieval can contaminate the illustration of inferred SCFs (Basu et al., 2018; O'Dell et al., 2018; Yu et al., 2018; Schuh et al., 2019)*".

Lines 66–67: "The COLA DA system consists of …, and the assimilated observations". I would not include the assimilated observations in the COLA DA system (by this logic, you should also include the meteorological driving data, the bottom-up fluxes, etc.).

Reply: Thanks for the comments and suggestion. We have deleted the "assimilated observations" in this context (Line 74). The meteorological and flux data are forcings that adding them to this sentence maybe too complicate. Thus, we prefer to describe them independently (Line 77).

Lines 76 and 91: "LETKF estimates SCF as evolving parameters" and "COLA treats SCFs as stationary parameters". I get what the authors mean, but for readers who are not familiar with the method, this could sound contradictory.

Reply: Thanks for the comments and suggestion. SCFs are evolving parameters between assimilation windows but stay stationary within the assimilation/observation window. The sentences are revised in Line 84, "*In the LETKF algorithm, the SCFs are treated as evolving parameters between assimilation windows*", and in Line 107, "*the SCFs are treated as stationary parameters within the observation window, meaning that they remain constant during model forecasting and are updated exclusively using the LETKF algorithm*".

Lines 76–77: "by augmenting it with the state vector CO2". I would rather say that "the state vector, which contains atmospheric CO2 concentrations, was augmented to also include the SCF parameters" or something to that effect.

Reply: Thanks for the comment and suggestion. The sentence is revised in Line 85, "*The state vector, which contains atmospheric CO_2 concentrations, is augmented to also include the SCFs parameters*".

Line 77: "the LETKF prefers a short assimilation window". I would argue that it depends on the application, and in some applications an ensemble Kalman Smoother with a longer assimilation window is more desirable.

Reply: Thanks for the comment and suggestion. We agree that it depends on the application. But, in this context, we meant for state data assimilation, for example in weather data assimilation and CO_2 state data assimilation. And after this sentence, we further argue that the flux parameter benefits from a long period of training (Line 88) that an ensemble Kalman Smoother with a longer assimilation window is more desirable.

Line 94: "There are two widely used a priori regularization approaches for EnKF-based carbon inversion systems." As previously mentioned, I believe this is misrepresenting the regularization approaches commonly used in inversions.

Reply: Thanks for the comment. Instead of referring them as traditional method, we just described these methods but not saying them as "traditional" and have revised the sentence to be more accurate, Line 111, "*However, for systems with longer assimilation windows, such as those spanning several weeks for SCF estimation, the temporal dynamics of the terrestrial ecosystem, including diurnal cycles and the transition from winter respiration to rapid spring growth, can introduce temporal lag and inaccuracies. Consequently, in CO₂ inversion systems with longer assimilation windows, the initial estimate of carbon fluxes heavily relies on a priori regularization techniques. For example, Feng et al. (2009) utilized the a priori information as the initial guess for their EnKF calculations, while Peters et al. (2007) employed the a priori to derive carbon flux variation modes and corresponding scaling factors, which are optimized through EnKF*".

Lines 98–99: "This approach omits useful information on the temporal dependency of SCFs". I do not understand the authors' point here—how does this omit information on the temporal dependence, given that the temporal variations are captured by the bottom-up fluxes?

Reply: Thanks for the comment. We are sorry that we misunderstand the method. Yes, the bottom-up fluxes provide the temporal variations. We have revised the description in Line 111.

Line 106: "we propose a new a priori regularization method to better follow the DA principle". The DA principle, as I see it, is to optimally combine a priori information with observations. I do not see how this new regularization method better follows the data assimilation principle, considering that it does not use an informative prior for the fluxes.

Reply: Thanks for the comment. This description is not accurate, and we revised it in Line 128, "*Based on the DA principle, we propose a novel a priori regularization method for COLA*".

Lines 127–128: "we define the SCF spatial gradient at a given grid as the SCF difference with its surrounding grids divided by the distance between them." Is this considering only grid points orthogonally adjacent, or also diagonally adjacent (I assume the former)? The gradient of a 2D field has two components, in this example x and y, but this is not mentioned anywhere. I suspect that the authors accounted for this given that Equations (5) and (8) have an overhead arrow for the gradient variables, but the meaning of the arrow is not mentioned anywhere. Additionally, the vertical bars are not defined—I assume they are used to denote the Euclidean norm?

Reply: Thanks for pointing out the missing information. Yes, we only considering the orthogonally adjacent grid points. We revised the sentence in Line 153, "*Numerically, we define the ∇f vector at a given grid as the SCFs difference with its orthogonally adjacent grids divided by the distance between them. Thus, there are two components in the ∇f along the longitude and latitude direction*".

In Equation 3 and 4, the vertical bar is the Euclidean norm. We have added the details in Line 207.

Lines 168–170: "A nature run is driven by the F_{OA} from Rödenbeck et al. (2014), F_{FE} from the Open-source Data Inventory of Anthropogenic CO2 emissions (ODIAC) (Oda et al., 2018), and the F_{IR} ..." Do these other flux components (ocean, fossil fuel, and fire) matter if they are identical in the truth and inversion experiments (considering that the atmospheric transport is perfect)?

Reply: Thanks for the comment. Since the transport is perfect, the other components do not matter.

Equations (7)–(9): Do t1 and t2 represent the actual time (as the text suggests), or are those time steps? If it is the latter, why was t2-t1 used rather than t2-t1+1? (The latter would correspond to division by the number of samples.)

Figure 3: The lower panel of Figure 3 shows that the RMSE and bias of EXP-NP are equal (1.36 GtC/yr). Normally in such a situation I would think that almost all of the RMSE can be explained by the bias, but this is not the case when looking at the lines in Figure 3. Do the RMSE and bias happen to be the same

here because the denominator uses N-1 rather than N (where N is the number of samples) in the definitions of RMSE and MB (see previous point)?

Reply: Thanks for pointing out the mistake. t1 and t2 represent the timestep. We have revised the Equation 2 and the description for the equation.

3. Technical corrections

Line 76: Remove parentheses around citation.

Reply: Thanks for pointing out the mistake. We have deleted the parentheses.

Line 77: "Similar to the other EnKF" -> "Similar to other EnKF variations"

Reply: Thanks for the suggestion. We have revised the corresponding sentence in Line 86.

Line 79: "long training period". Do the authors mean "a long assimilation window"? "Training" is not defined in the manuscript.

Reply: Thanks for the comment. "Training" may mislead readers. We replaced the "long training period" to "long observation window".

Line 97: "and the subscripts b, p, and t" -> "and the superscripts b and p and subscript t"

Reply: Thanks for pointing out the mistake. Based on your previous major comments, we have deleted the corresponding descriptions.

Figures 6 and 7: Change "Kg" to "kg".

Reply: Thanks for the suggestion. We have changed "Kg" to "kg" in Figure 3.

Line 232: "show an increase in RMSE of FTA concerning EXP-NP" Do the authors mean "compared with" instead of "concerning"?

Reply: Thanks for the suggestion. Yes, it means "compared with". Since we have made a large revision on the results section. The corresponding sentence was deleted. And we are sorry about this.

As mentioned there could be other issues that I have not raised here, which can be properly assessed only after the authors have included the missing information and clarified the manuscript.

Reply: Thanks for the careful readings. We have careful revised other sentences and hope that we have addressed your major concerns.

Table 1: Assimilation experiments setup.

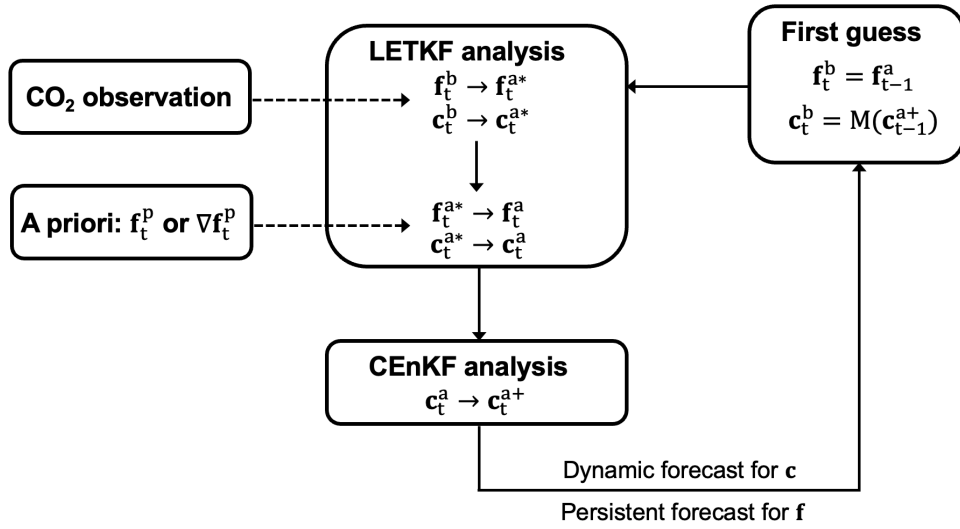| Experiment	EXP-GOI	EXP-FOI	EXP-OI	EXP-GI	EXP-FI
A priori	$\nabla \mathbf{f}$	$\mathbf{f}$		$\nabla \mathbf{f}$	$\mathbf{f}$
Observation	OCO-2+In-situ	OCO-2+In-situ	OCO-2+In-situ	In-situ	In-situ

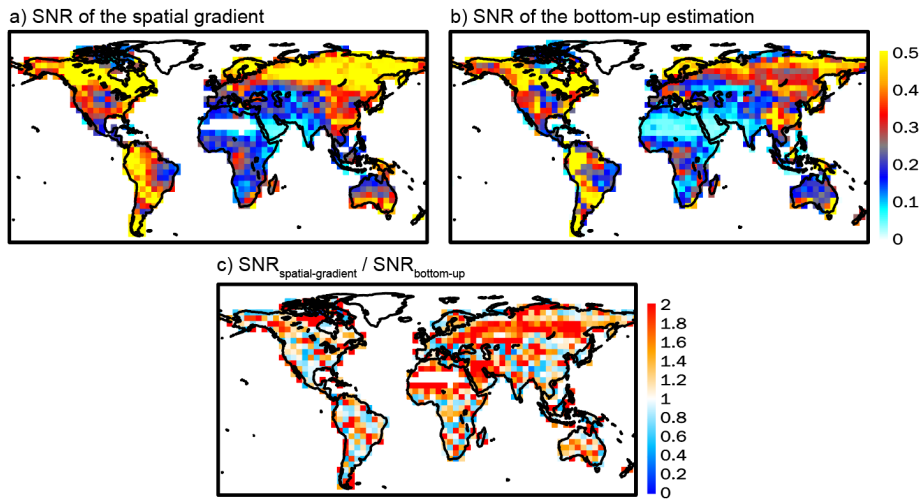**Figure 1: The assimilation cycle of the COLA system, illustrating how and where the a priori is assimilated.**

Figure 2: a) The annual mean signal-to-noise ratio pattern of the a priori F_{TA} spatial gradient. b) The same as a) but for the a priori F_{TA}. c) Annual mean signal-to-noise ratio of the spatial gradient divided by the signal-to-noise ratio of the F_{TA}.
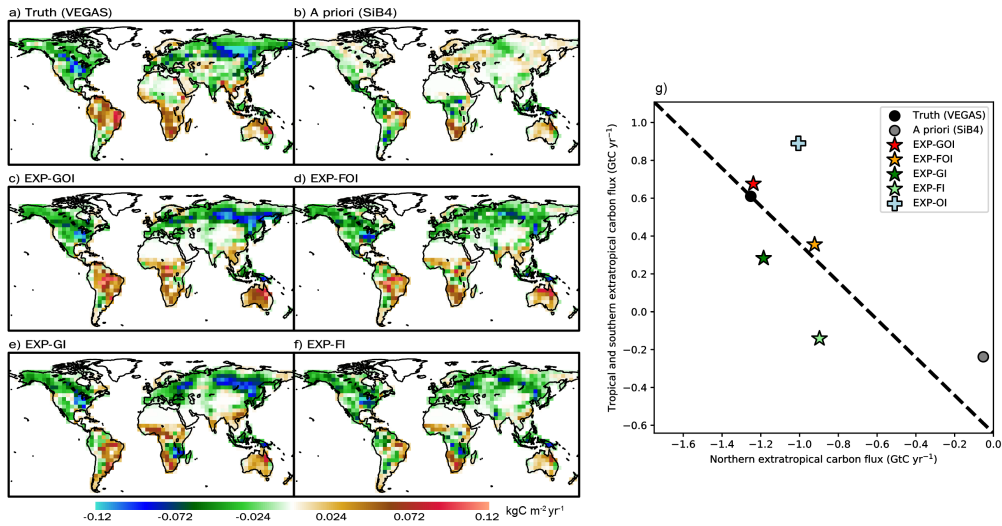
Figure 3: The annual mean F_{TA} in 2015 of a) the truth, b) the a priori, c) EXP-GOI, d) EXP-FOI, e) EXP-GI, and f) EXP-FI. g) Comparison of the annual total F_{TA} in the northern extratropical area and the tropical and southern extratropical areas. Different scatters denote the truth, the a priori, and the different experiments. The dashed black line denotes the global carbon budget.
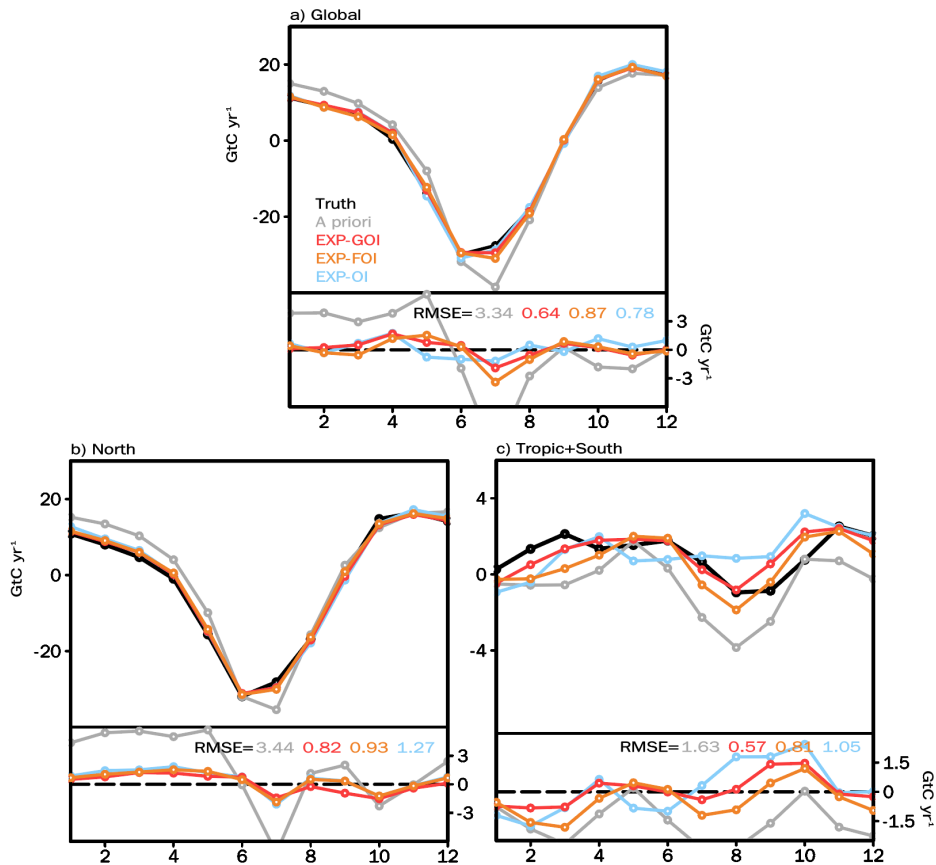

Figure 4: The top figures in each subplot are the a) global, b) northern extratropical, and c) tropical and southern extratropical seasonal cycles of F_{TA} in the truth (black), the a priori (gray), EXP-GOI (red), EXP-FOI (orange), and EXP-OI (sky blue). The bottom figures in each subplot are the global total difference of the a priori and the three assimilation experiments compared to the truth. The annual mean RMSEs of the a priori F_{TA} and the three experiments are denoted at the upper right corner of the bottom figures.
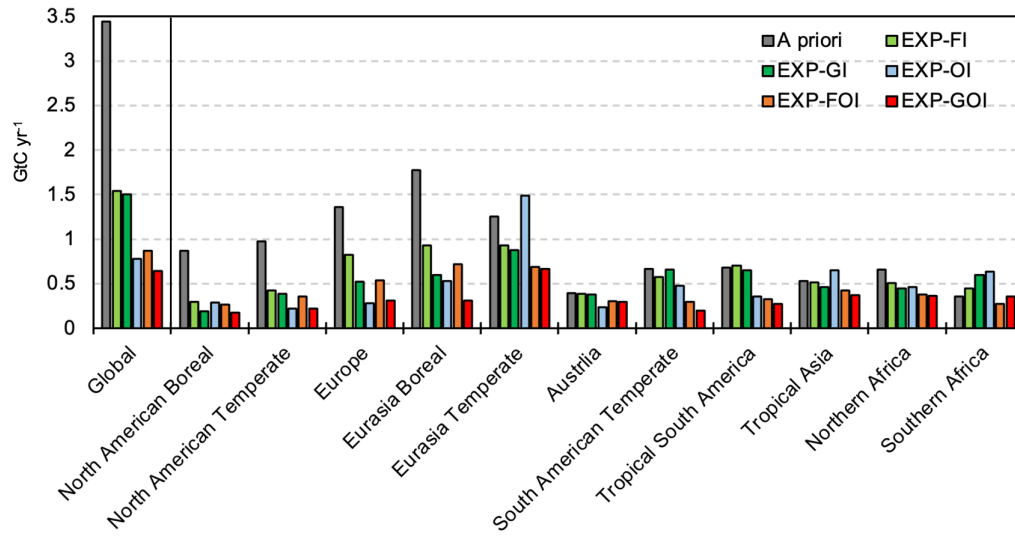
Figure 5: The regional RMSE of the a priori and the five assimilation experiments compared to the truth. The regions are defined by the OCO2MIP.