Thanks for the careful reading and insightful comments. We have carefully revised the manuscript accordingly.

The authors investigate the benefit of constraining the CO2 mass when simultaneously estimating the CO2 (state variables) and the surface carbon fluxes (parameters) with the LETKF in an idealized set-up. The science is valid and benefits the corresponding research field. However, major improvements are needed in terms of language and the discussion of the results. I listed the major issues of this paper below.

Language:
Unfortunately the English is very poor, which makes it is sometimes difficult to decipher what the authors mean. Some examples:
- line 66: “The system replaces the GCM …”. What is the GCM replaced with? Is it with GEOS-Chem? This is not clear from the sentence.
- line line 207: “To get the prior ensemble ….” What does it mean: “at 1 October 2014 within 30 days”?
- line 264: “However, the SC amplitude and the phase are reinvestigated…” I don’t understand the word reinvestigated in this context.
Many articles (“the” and “a/an”) are either missing or are there where they shouldn’t be.
Response: Thanks for your careful reading. We revised the manuscript in terms of grammar and words. In addition, the manuscript was polished by a native speaker from AJE (America Journal Expert).

In my opinion there are too many abbreviations introduced, which makes it hard to read the paper. Perhaps including a table with all abbreviations would help.
Response: Thanks for the suggestion. We have deleted some abbreviations and added a table to the appendix section (Table. A3).

In the abstract, line 23, the authors state that they introduce a Constrained Ensemble Kalman Filter. However, if I am not mistaken, this was introduced by Pan and Wood (2006).
Response: Thanks for pointing out this mistake. We replaced the word 'introduce' with 'apply'.

Provided information:
It was necessary to read Liu et al 2019 to understand the current manuscript. I feel that the authors did not optimize the selection of information given. For example, the authors chose to write down the equations for the LETKF, but did not explain what an “observation window is”. In my opinion the authors have 2 choices:
1) Let the readers explicitly know that this is a follow up paper of Liu et al 2019, so that they know they should read Liu et al 2019 first. In this case, a lot of the paper up to section 2.3 should be shortened (and also section 3).
2) Make sure that this paper is understandable without Liu et al 2019. Specific examples of information that needs to be added are:
- explaining the “running in place” principle. This can be very short as is for instance done in the abstract of Liu et al 2019.

Response: Thanks for the suggestion. We prefer the second choice, and we also guide the readers to read Liu et al. (2019) for more information. We added a flowchart figure to describe the whole COLA system (Fig. 1). It described how AW and OW worked and how the two steps of EnKF were applied. We believe this figure could help readers further understand the COLA system. In addition, for the methodology section, we further explained the method related to the short assimilation window and long observation window (Line 140).

- The authors are estimating the SCF, but the results show fields of FTA. Do they calculate FTA from SCF? Clarify! Also, what is E_t in equation 12) (from Liu et al 2019 I deduce it is the time average)?

Response: Thanks for pointing out this. The FTA is part of SCF (SCF=FTA+FOA+FFE, Line 110). In the OSSE set-up, we only focus on estimating the FTA, while FOA and FFE are directly prescribed as 'background' fluxes (Line 210) that are the same in nature and assimilation runs (Table. 1). And E_t is the time average. We have added more details to clarify the equation (Line 279).

- Line 66: “The system replaces …” Elaborate on the reasons why the GCM model is replaced. The GEOS-Chem model has not been introduced, so the reader does not know that it does not include an estimation of transport uncertainties related to the meteorological field. Again, this is well explained in the abstract/introduction of Liu et al.2019. Also, the authors should shortly discuss the implications of assuming perfect meteorological fields. Is it a reasonable assumption, i.e. are the errors small in comparison to CO2 and the SCF in reality? Or do the authors expect it should not impact the estimation of SCF a lot? Or is this assumption made to isolate the effect of LETKF_C on the SCF?

Response: Thanks for the suggestion. We briefly add some discussion on this (Line 68). Although the online modeling of the atmospheric dynamic could estimate the transport uncertainties. The quality of the GCM modeled atmospheric transport is usually poor than the atmospheric reanalysis data (e.g., MERRA2). And the computational cost is very expensive. Thus, using the offline ATM instead of the online GCM is an appropriate approach. And this approach does not include the estimation of transport uncertainties related to the meteorological field, which will lead to additional errors for SCF estimation in reality. This assumption is the commonly used in CO_2 inversion studies. The impact is assumed small but remains to be validated in the future. We can include the meteorological field uncertainties by driving the ATM using different reanalysis products for different ensemble members. Such a capacity is under development.

- The introduction of COLA is very confusing. Did I understand correctly that COLA is the name of the system which uses the GEOS-Chem model as the ATM and LETKF_C (without or without CEnKF) as the DA algorithm? Please clarify in the manuscript.

Response: Thanks for pointing out this mistake. LETKF_C is the system name but not the DA algorithm (please see Liu et al. 2019, section 2). Based on the LETKF_C system, the COLA system is developed with an improved framework (inflation schemes, CEnKF, etc.). We add more details in the introduction section (Line 80).

- How is the inter-annual variation calculated?

Response: Thanks for the comment. It is calculated using the 12-month moving average method

(Line 306).

- Many figure captions miss information (for example I assume the shaded region in Figure 4a is spread, but it is not stated)

Response: Thanks for the comment. We have carefully revised the figure captions.

- The RMSER is introduced and is even discussed as if the RMSER is presented, but none of the Figures actually show the RMSER. They seem to show the difference compared to the truth.

Response: Thanks for pointing out this mistake. We have added a table (Table A2) summarizing the RMSER of EXP-L, EXP-LC, and EXP-LC.

Science:

I am concerned that the authors chose to constrain the CO2 mass on the ensemble mean only. They state in line 140: "We further simplified the method by constraining only the ensemble mean state, which significantly reduced the computational cost without influencing the performance". However, the authors provide no evidence that the performance is not influenced. By constraining only the mean, each ensemble member is free to violate the mass conservation. However, as Pan and Wood 2006 point out, adding the mass conservation constraint essentially adds physical information to the DA process. By not enforcing the mass constraint on each member, a lot of this physical information is lost. Also, as the authors point out in line 75, "the impact of mass gain or loss could last for a long time". I would like to add that the impact is not necessarily linear in time, nor symmetric. As a result, a CO2 forecast could be very different when constraining each member as opposed to only the mean. I think that constraining the mean is still helpful because it will likely (though not guaranteed!) reduce the imbalance in each member anyway, but it remains an "ad hoc" solution forced by computational restrictions. I therefore think that the authors should do one of the following:

Do an experiment with all members constrained. It would really be great if this could be done, but I am not aware of the computational limitations the authors have. If it is not possible, perhaps a set of experiments EXP-L, EXP-LC and EXP_LC2 (where EXP_LC2 corresponds to constraining each member) is feasible with a smaller ensemble size?

Elaborate on this point. Explain that it is not possible to constrain each member and try to justify the choice of only constraining the mean and discuss the possible drawbacks. In Figure 11, show also the ensemble imbalance spread for EXP-LC, not only EXP-L.

Response: Thanks for the comment and suggestion. There are two reasons that why we decide to constrain the ensemble mean only. 1) The RTPS inflation step will destroy the balance within each ensemble member. 2) Computationally expensive when run at high spatial resolution. At the beginning, we have applied the CEnKF to constrain each ensemble member. And the computational cost of CEnKF was ~10 seconds/cycle. Such cost is comparable to the cost of LETKF and thus acceptable (Table A1). But the time cost would increase much when we run at 2×2.5 resolution (~15 minutes/cycle). Since more and more CO_2 observations will be available in the future, higher spatial resolution CO_2 data assimilation is urgently needed. Thus, we simplified the method to constrain only the ensemble mean that we could run at higher spatial resolution.

Indeed, we need to clarify the possible effects. We conduct a new experiment (EXP-LCE) that constrain each member. We summarized the performance of EXP-LC and EXP-LCE in table A2 in terms of RMSER. Our conclusion is that the performance of EXP-LC is similar to EXP-LCE (Line

294). We hope the new experiment could address the concern.

A large portion of the results is dedicated to evaluating EXP-L with respect to the prior. However, this is what was covered in Liu et al 2019. The current paper should focus on the comparison between EXP-L and EXP-LC with maybe the prior as a helpful benchmark. I therefore think that the abstract sentence on line 28 “At the seasonal scale …” should be deleted. Also because it is not clear to the reader that the “improved system” is referring to EXP-L, not EXP-LC. This confusion is also very much present in the conclusion, where I also think the improvement of EXP-L over the prior should not be highlighted as a result. I therefore also suggest that the authors merge section 4.1 and 4.2 and basically make all plots with prior, EXP-L and EXP-LC.
Response: Thanks for the comments. First, I guess the reviewer has misread Figure 5 and 6 (now Fig. 6, 7). We presented the EXP-LC but not the EXP-L compared with truth and a priori. As in Liu et al. 2019, it only focuses on the global seasonal cycle, while the region seasonal cycle and carbon budget has not been discussed. For CO_2 inversion studies, the estimation at global scale is much more precise compared with the regional scale. And for those who are interested in CO_2 inversion, they would like to see a comprehensive analysis from global to regional. Thus, we present Figure 4~7 first that the readers could have a first impression of the COLA system. We organized the article in this form so that both CO_2 inversion researchers and data assimilation researchers may be interested in it. So, we prefer to retain the seasonal scale analysis in both the abstract and main text.

If the reason the authors did not include EXP-LC in Figures 4-7 is that the difference between EXP-L and EXP-LC is not visible, this should be explicitly stated. Perhaps the authors meant to communicate this on line 302, but this was not clear to me. A likely reason for the lack of difference is that the SCF are updated using the covariances between CO2 and SCF, which are probably not as strongly impacted by the mass constraint as the ensemble mean. One would see a much greater effect on a CO2 forecast.
Response: Thanks for the comments. We further explained this at line 271. Yes, the difference between EXP-L and EXP-LC is not visible, and therefore we only presented EXP-LC. Another reason for the lack of difference is that LETKF is accurate in estimating the large seasonal cycle amplitude. However, the mass loss of ~0.2 GtC/year is much smaller than the seasonal cycle amplitude of ~40 GtC/year. While averaging to the annual mean, the ~0.2 GtC/year mass loss becomes comparable to the ~1.2 GtC/year carbon sink, and the effect of CEnKF will come in at the annual scale. The impact of CEnKF on the CO2 forecast is directly visible (Fig. 12). But such an effect will reduce after a long time (Fig. A1). And on the contrary, the impact on the flux will stand out after a long time.

The RMSER is introduced but not presented. I feel the RMSER (both as a function of time, averaged over space (Figure 4b), and as function of space averaged over time (Figure 7) would be an efficient and effective way to present results. I would be interested in both the RMSER of EXP-L with respect to the truth, and the RMSER of EXP-LC with respect to EXP-L.
Response: Thanks for the comments. We have shown the RMSE in Figure 5~7. And the RMSER is calculated based on the RMSE. But we did not explicitly show the RMSER. Thus, we added Table. A2 to show the RMSER for each region.

More results could be shown. For example, the inflation scheme for C02 is different than in Liu et al 2019. It would be nice to show the spread and RMSE (or spread skill ratio) of both CO2 and the SCF throughout the experiment period, including the spinup. Also, more could be said and shown about in effect the mass constraints have on the SCF increments and covariances between CO2 and the SCF. I would also be curious to look at the background RMSER (EXP-L with respect to EXP-LC) of the background CO2, not only the increments.

Response: Thanks for the comments and interest. I agree that more results could be showed. But this manuscript is focused on the CEnKF impact and the overall performance of COLA. We plan to discuss some of them in future works.

Plots:

Figure 7 and especially Figure 10 are too hard to read. Either make the plots bigger or show the RMSER with a colorbar that is centered aroud zero, so that it is easy to spot the improvement regions.

Response: Thanks for the suggestions. We have made the plot bigger (Fig. 8). For Figure 10 (now Figure 11), because there are many grid points that the values are very small (near zero). The relative error reduction could be very large that making the plot very noise (Fig. R1). So, we prefer to use the original form.

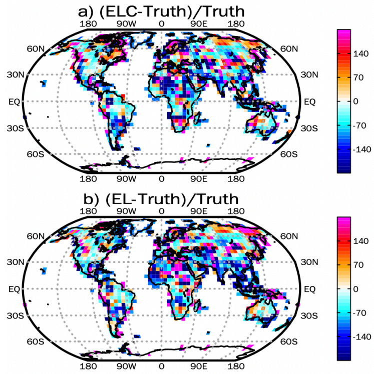

Figure R1: The percentage of the difference of EXP-LC and EXP-L compared with the truth.