Response to reviewer 1:

The authors partially adressed my concerns. There was a large improvement in terms of readibility of the paper. I still have two major concerns, which I think the authors should consider, as well as several minor ones. I acknowdledge that the second major concern is somewhat subjective.

Response: Many thanks for the very careful reading and the critical comments. We have revised the paper correspondingly.

Major comments

- The authors claim that the RMSER for constraining the mean and contraining each member are similar, but from Table A2 I have to disagree. The differences between EXP-LC and EXP-LCE are at least the same order of magnitude as the differences with EXP-L. So if the authors claim that the differences between EXP-LC and EXP-L are significant, then so are the differences between EXP-LC and EXP-LCE. This definitely needs attention!

Response: Many thanks for pointing out this issue that may be an overstatement. I agree that the differences between EXP-LC and EXP-L are not significant and the main advantage of the CEnKF is the global total FTA (the reviewer also points out in the minor comments that " *Based on the results I think the main advantage of the imblance constraints is the global total FTA. When zooming into regions as in Figure 10, it is not convincing that EXP-LC is better than EXP-L. Although the authors are not writing something that is untrue, if feel that this text is misleading. I suggest that the authors write that EXP-LC does not convincingly outperform EXP-L on a regional level.*"). We rephrased the corresponding statements. But we still think that EXP-LC shows slight, even though not significant, improvement at the regional scale. We think they are meaningful and worth noting. The reasons are as follows:

1) In terms of the annual mean spatial pattern (Fig. 11, R1), we find some improved regional features, such as EXP-LC successfully captures the carbon source at the side and the carbon sink at the center (the red rectangular region in Fig. R1) (**line 384**). I admit this may be 'cherry picking'. But I think it is meaningful, at least.

2) In terms of regional RMSER, we replaced the RMSER table with a bar plot (Fig. A2), and it indicates that all the experiments show similar RMSER over the northern hemisphere regions. But we can see clear differences over the tropical and southern hemisphere regions where there are much fewer surface observations. And we find that EXP-LC is better than EXP-L over all the tropical and Southern hemisphere regions, which indicates that the CEnKF can potentially improve the performance over the poorly observed regions (**line 300**).

3) In terms of regional annual total FTA (Fig. 10), we can see a similar slight improvement over the tropical and Southern hemisphere regions. The EXP-LC is better than EXP-L over most of the tropical and Southern hemisphere regions except the South American Temperate region (**line 359**).

Thus, we claim that EXP-LC is slightly better than EXP-L over the poorly observed tropical and southern hemisphere regions.

We absolutely agree with the reviewer that "*The differences between EXP-LC and EXP-LCE are at least the same order of magnitude as the differences with EXP-L*". And the differences also appear over the tropical and southern hemisphere regions. This is what we missed in the last round of revision. We added some statements (**line 300**). The EXP-LC shows larger RMSER compared with EXP-LCE over Australia, northern tropical South America, and southern Africa regions. And EXP-LCE shows larger RMSER compared with EXP-LC over South America Template and northern tropical Asia regions. Overall, the EXP-LC is not significantly worse or better than EXP-LCE over the tropical and southern hemisphere regions in terms of RMSER, which further proves that the simplified CEnKF (constrain ensemble mean only) does not degrade the performance compared with the original CEnKF (constrain each ensemble member). We add more statements to clarify those differences.

In conclusion, we did find some evidence that EXP-LC is slightly but not significantly better than EXP-L. Those evidence are meaningful, but we can not prove that they are significant. Thus, we clarified the corresponding statements. And we find that there are differences between EXP-LC and EXP-LCE in terms of RMSER. But these differences do not prove that EXP-LC is worse than EXP-LCE.

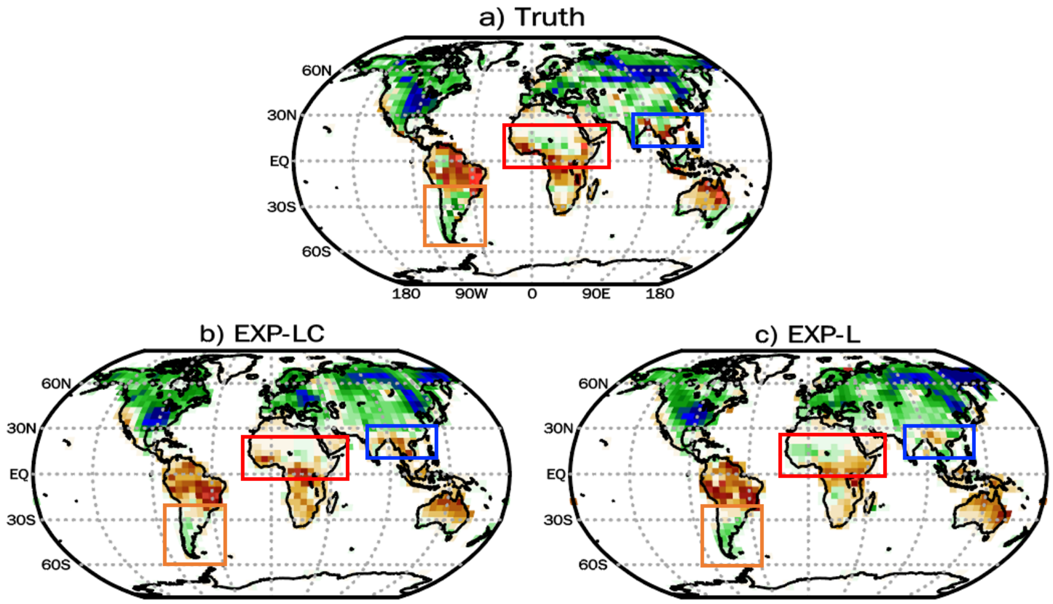

Figure R1: The spatial distribution of FTA for the truth (a), EXP-LC (b), and EXP-L (c) averaged from January 2015 to December 2017.

- The authors call the COLA system an "improved system". I assume the authors mean improved with resepct to the LETKF_C system in Liu etal 2019. However in section 4.1 the authors don't show the improvement with respect to the LETKF_C system, but with respect to the prior. I already addressed this point in my previous review. The fact that the LETKF_C system is working has already been covered in Liu etal 2019. Correct me if I am wrong, but the two subsequent paragraphs in the conclusion starting line 411 already applies to LETKF_C and therefore have been covered in Lius etal 2019. It is not a new conclusion. The purpose of this paper should be to support the claim that the system is an improvement over the LETKF_C (and with that also the prior), so COLA should be compared to LETKF_C. If the difference between LETKF_C and COLA are too small to show, then the authors only need to write a sentence that COLA performs as well as LETKF_C on seasonal cycle and interannual variation. I think that the message of this paper is that applying the imbalance constraints reduces the global bias of SCF. This message it definitely worth communicating and it does not require a long paper. I think a more concise paper is favourable in this case.

Response: Thanks for the comment. We build the COLA system based on the LETKF_C system. And, yes, we call the COLA system an "improved system" as compared with LETKF_C. I must admit that use the word 'improved' can mislead the readers. Thus, we deleted the word 'improved'. Besides the CEnKF and RTPS scheme implemented in COLA, there are many specific changes from LETKF_C to COLA. We update the GEOS-Chem model from version 10.01 to 13.0.2. We made a great improvement on the coding structure, such as: 1) using the 'spack' software on controlling the computing environment as suggested by the GEOS-Chem team. 2) producing the ensemble simulations by running a single GEOS-Chem instead of GEOS-Chem ensembles. 3) easy to switch the meteorology fields and the a priori fluxes.

Thus, since there are so many changes, it would be very hard and not practical for us to go back to the original LETKF_C system and conduct another experiment. In this paper, the EXP-L is similar to the original LETKF_C configuration. So, we treated the EXP-L as a baseline (a proxy of LETKF_C) and showed the improvements compared with it. As the reviewer pointed out in the last round of revision, we have added some statements to clarify that the difference between EXP-LC and EXP-L is not visable at the seasonal scale and the improvements showed up at the annual scale (annual total and interannual variation, Fig. 9) (**line 275**).

In the conclusion and discussion section starting from **line 411** (**now line 433**), these are discussions but not the conclusion. We discussed the window length compared with traditional methods using a very long observation window (3 months to 1 year). Using a very long window is a standard configuration in CO_2 inversion studies. We emphasized that the ensemble-based methods using a short window with the persistence forecast hypothesis (the basic EnKF parameter DA configuration) can also yield accurate results. This discussion is very important for the CO_2 inversion community and has not been discussed in Liu et al. 2019. We made a large revision to this discussion to make it clearer.

Finally, we want to emphasize that Liu et al. 2019 is a preliminary attempt to prove the method works. However, it only showed the results of the global seasonal cycle and the global spatial pattern. While some critical analyses like the global/regional annual total budget and the regional seasonal cycle are missed. We extend the analysis in this paper to show the robustness of COLA.

Minor comments

- line 37: Recently, many countries (e.g. Asian .... and South American countries) announced ... --> Recently, many countries in for instance Asia, ... and South America announced ...
- line 57: from Gaussian distribution --> from a Gaussian distribution & with long AW -- > with a long AW

Response: Thanks for pointing out these mistakes. We have revised them.

- line 68: Explain briefly the concept of an observation window (does it mean that observations are assimilated multiple times?)

Response: Thanks for the suggestion. We have added some explanations (**line 69**).

- line 69: cost is very expensive -- > cost is very high
- line 94: present --> presents

- line 117: carbon data assimilations --> carbon data assimilation
- line 119: generate the analysis --> generate an analysis
- line 122: The y_k^b=h(x_k^b) --> y_k^b=h(x_k^b) (remove "The")
- line 141: uses an unique --> uses a unique
- line 142: in daily bases --> on a daily basis
- line 152: we choose only --> we choose to only & on ensemble mean --> on the ensemble mean & instead of each ensemble member --> insteand of on each ensemble member

Response: Thanks for pointing out these mistakes. We have revised them.

- line 183: "negative ensemble variance" Variance cannot be negative. Do the authors mean reduction of ensemble variance?

Response: Thanks for pointing out this mistake. We have revised it. Yes, we mean the reduction of ensemble variance.

- line 205: superscript ps --> superscript p
- line 256: 3 OSSE Results --> 4 OSSE Results

Response: Thanks for pointing out these mistakes. We have revised them.

- Table 1: This table makes it more confusing for me, instead of less. For example, the assimilation window, Observation window, Ensemble member (which should be Ensemble size), FTA, FOA and FFE apply to all assimilation runs, right? Use lines to make clear that these number hold for all 3 experments and be consistent with the spacing (FOA and FFE are under EXP-LC and the other value are under EXP-L)

Response: Thanks for the comment. We have added lines to make the table clearer. Yes, the assimilation window, Observation window, Ensemble size, FTA, FOA, and FFE apply to all assimilation runs.

- equations 12 and 13: Are the RMSE_reg^a and RMSER_reg^a both a function of space? Because the authors present the RMSER as a number, so I assume that is some averaging involved, which is not mentioned. How is the averaging done? First averaging RMSE_reg^p and RMSE_regâ over gridpoints and then calculating the RMSER, or is the spatial averaging done in the end?

Response: Thanks for the comment. We have further modified the formula to make it clearer. Yes, both $\mathrm{RMSE}^{\mathrm{a}}_{\mathrm{reg}}$ and $\mathrm{RMSER}^{\mathrm{a}}_{\mathrm{reg}}$ are a function of space (subscript reg). The spatial averaging is done at the beginning instead of in the end. The $\mathrm{RMSE}^{\mathrm{a}}_{\mathrm{reg}}$ is defined for each continental region (regions are defined in Fig. 6 and 7). Thus, before calculating the $\mathrm{RMSE}^{\mathrm{a}}_{\mathrm{reg}}$, we calculate the regional total $\mathrm{FTA}_{\mathrm{reg}}(\mathrm{T})$ at each time (T). Finally, we calculate the RMSE based on the time series of $\mathrm{FTA}^{\mathrm{a}}_{\mathrm{reg}}(\mathrm{T})$ and $\mathrm{FTA}^{\mathrm{t}}_{\mathrm{reg}}(\mathrm{T})$. So, the RMSE and RMSER is a number instead of a series.

- Figure 5: The authors claim that "The ensemble mean initial SCF and CO2 conditions are significantly larger than the truth", which is why spinnup is needed. I am not objecting against spinnup, but the spinnup period does not look much different from the rest of the graph in Figure 5. Can the authors comment on that? Also, in caption of Figure 5 and 6 it says that the RMSE is shown based on equation 13, but equation 13 is the RMSER.

Response: Thanks for the comment. We add a figure to the appendix section to show the IC (Fig.

A1) and add more descriptions on the IC (**line 226**). I agree that the difference at the IC in Figure 5 is not significant. Because Fig 5 presents the global total flux. If we look at the spatial pattern of SCF at the IC (Fig. A1). The difference is large over the Eurasia/North America boreal regions. Moreover, if we look at Fig 12g, it shows a clear negative imbalance at the beginning that indirectly shows that the ensemble IC is biased. And thanks for point out the mistake that the equation should be 12 instead of 13.

- line 343: "For EXP-LC without ..." I don't understand this sentences. Perhaps the authors meant to communicate that the anntual total FTA is increased with only 0.06 GtC with resepct to the truth?
Response: Thanks for the comment. Yes, we meant that the annual total FTA is increased with only 0.06 GtC with resepct to the truth. We revised the statement (**line 355**).

- line 346: "Regionally the performance ..." Based on the results I think the main advantage of the imblance constraints is the global total FTA. When zooming into regions as in Figure 10, it is not convincing that EXP-LC is better than EXP-L. Although the authors are not writing something that is untrue, if feel that this text is misleading. I suggest that the authors write that EXP-LC does not convincingly outperform EXP-L on a regional level.
Response: Thanks for the suggestion. We revised this statement (**line 359**). As the reviewer point out in the first major concern, we aggree that the differences are not significant. But some slight improvements are worth noting.

- Figure 9: Is it necessary to show both bias and annual total FTA? Is there information in the bias that we cannot infer from the annual total FTA?
Response: Thanks for the comment. We think it is necessary to show both bias and annual total FTA. The bias could be compared with the imbalance. We show both the bias and imbalance that the readers can clear see the difference (the bias is the difference between analysis and truth, the imbalance is the difference between the first guess and the analysis) and connection (EXP-LC: small bias because of no imbance problem; EXP-L: large bias because of large imbalance) between bias and imbalance.

- line 369: over the southern China --> over southern China & I don't now what is meant with "reinvestigated" in this context.
Response: Thanks for the comment. The word 'southern China' is not precise. We replaced it with 'Indochina'. We meant that the carbon source over 'Indochina' and the carbon sink over southern South America are captured in EXP-L and EXP-LC (blue and orange rectangular regions in Fig. R1).

- line 372: "Even though ..." I agree that the difference between EXP-LC and EXP-L is not significant, so the claim in the rest of the sentence is confusing to me and probably an over statement.
Response: Thanks for the comment. We revised the statement (**line 385**).

- line 381: "The spatial patterns of the LETKF ..." From the snapshot we can indeed see that in this case the increments of LETKF and CEnKF generally have opposite sign. It would be nice to back up this statement with a more statistically signifcant varification metric, such as the correlation.

Response: Thanks for the useful suggestion. We draw a new plot (Fig. 13) We calculate the spatial correlation between the LETKF increment and the CEnKF increment. And we linked this correlation with the LETKF imbalance. We find the magnitude of the increment correlation has a moderate relationship with the absolute global LETKF mass imbalance (**line 401**).

- line 400: "The COLA system shows improved performances" improved compared to what? We saw no evidence that it is improved with respect to LETKF_C.

Response: Thanks for the comment. We deleted the word 'improved' as in the second major concern.

- line 403: LETKF --> the LETKF & efficiently --> effectively

Response: Thanks for pointing out this mistake. We have revised them.

- line 405: "but improved the LETKF estimation". Did the authors provide evidence to support this claim? What is for example the average (RMSE^EXP-L- RMSE^EXP-LC)/RMSE^EXP-L?
- line 405: "Moreover, the ..." Again, I am not convinced of this claim. Perhaps the authors should be more humble with the wording.

Response: Thanks for the comment. They are overstatements and not precise. We revised that summary pragraph (**line 427**).

Response to reviewer 2:

The revision answered almost all of my questions. It should be accepted for publication just after minor corrections.

Response: Many thanks for the very careful reading and the critical comments, we have revised the paper correspondingly.

1. Line 27, Page 1: 'we show that this system can accurately track the annual mean SCF from global to grid-point scale'

The statement is too strong. For example, Biases over Eurasia boreal are still significant

Response: Thanks for the suggestion. We made some changes on the statements in the abstract. This sentence was deleted.

2. Line 55, Page: '...compromising the
sparse and unevenly distributed...'

Not sure what it means

Response: Thanks for the comment. We revised this sentence (**line 53**). '*Thus, to compromise the sparse and unevenly distributed feature of the global CO2 observation network, most top-down systems do not localize the observations and set a very long assimilation window (AW) that ranges from several months to one year*'.

3. Line 124, Page 4: '..with the observation operator **h**'

More details about h would be helpful.

Response: Thanks for the suggestion. We add some details about **h (line 125)**.

4. Line 141, Page 5: '...uses an unique setting of LETKF with short AW of 1 day and a long observation window (OW) of 7 days...'

It is interesting to know how the authors calculate the uncertainty for annual total flux (i.e., whether the temporal correlation has been taken into account.

Response: Thanks for the comment and interest. We did not consider the temporal correlation because we did not explicitly assign the temporal correlation in the flux ensembles. And we additively inflate the ensembles based on the variability of priori fluxes. But there are definetely correlations between adjacent assimilation windows since we use the persistent forecast model. Thus, to calculate the annual total uncertainty, we objectively assume that there is no correlation between each month and calculate the annual total uncertainty based on the sum of monthly flux uncertainty. We believe that there are more accurate methods based on the temporal correlation of flux ensembles. And the work is under development and will be discussed in our real observation DA papers.

5. Line 162, Page 6: '...where **h** is the linear “observation” operator..'

Using 'h' for the observation operator again can cause confusion with the one defined in Eqs.1-4

Response: Thanks for pointing out this problem that may confuse readers. We replace $\mathbf{h}$ with $\mathbf{h}^{'}$.

6. Line 174, Page 6: 'The grid with a larger ensemble spread will likely give more mass constraints.'

Should it should 'get' not 'give' ?

Response: Thanks for point out this mistake. We replace the word 'give' as 'get'.

7. Line 274, Page 9: 'SC amplitude...',

please define SC (I assume it be Seasonal Cycle)

Response: Thanks for point out this mistake. We revised this.

8. '...the SC phase shows a one-month lag, ...'

I would like to see what is the cause for the one-month phase lag.

Response: Thanks for the comment and interest. Such temporal lag is not well understood. We think this is likely because of the sparse observations over the tropical South America.

9. Figure 6: The correlation between true and posterior IAV should be shown in the plots.

Response: Thanks for this useful suggestion. We add the correlation values to Fig 6 and 7.