Reply to Reviewer #2

The manuscript written by Peiro et al., (2021) is an update of the study made by Crowell et al., (2019), et al., (2019), who assessed the annual and monthly ensemble mean flux derived by nine different global inversions using the assimilation of OCO-2 satellite data (version 7) and in situ data for the period 2015-2016. Helene et al., 2019 used the findings found in Crowell et al. (2019) and compared them with the ensemble median flux of 10 different global inversions based on the assimilation of OCO-2 (version 9) land nadir and land glint (LNLG) data. Further, the authors also analysed fluxes for a more extended period 2015-2018, using both in situ and OCO-2 satellite data separately. In general, the manuscript is well-written, and the discussion and conclusions are interesting. However, some parts in the methodology and results sections need revision and clarifications. I would recommend this paper for publication once the authors have addressed all the questions described below.

We are grateful for the Reviewer comments and for taking the time to review our manuscript. We answered below the comments with information on page and line numbers that have been changed in the manuscript when necessary.

General comments:

1) Line 15. “However, the lack of data in the tropics limits our conclusions and the estimation of carbon emissions over tropical Africa require further analysis”. It would be good If the author can provide spatial maps of LNLG OCO-2 across the world for the studying years to see how bad the coverage is in Tropics. (Maybe show these maps in the appendix of the manuscript).
Thank you for this comment. We took it in consideration and included a map with LNLG OCO-2 10s retrievals in the annex and added a sentence line 268 page 14 “Figure. A1 in the appendix represents the locations of the OCO-2 10s retrievals LNLG for the period of study 2015-2018. We can see that the posterior flux estimates are constrained with OCO-2 LNLG observations particularly present in the Northern Hemisphere and with a lower number of observations over the tropics.”

2) Line 137. “This bias correction in v9 allows a more uniformly distribution of XCO_2 over regions of interest, decreasing the standard deviation to 0.74 ppm compared to v8, which was of 1.35 ppm.” Please indicate that this reduction was found in TCCON Lauder, New Zealand (Fig.13).
We considered this remark and modified the sentence accordingly, line 147, page 7:
“This bias correction in v9 allows a more uniformly distribution of XCO_2 over regions of interest, decreasing the standard deviation over the TCCON Lauder (New Zealand) site, for instance, to 0.74 ppm compared to v8 which was of 1.35 ppm.”

3) Line 147-151 Interestingly, the authors accounted for correlations between the model-data mismatch error for the 10 second average of OCO2 soundings. However, the description of how it was calculated is not clear. It would be great if the author could provide more details in this section of the manuscript. Line 159. “Both resulting uncertainty is then passed through the formula to account for error correlation.” Which formula? It is hard to understand what the author did to calculate the observation uncertainties. Please provide more details in this section.
We added more information on the error calculation on the 10s average to be more specified of what have been done in reference to the study of Baker et al., 2021. The formula used to account for error correlation was also added in the this section. We consequently changed line 165, page 7 to :
“Since it is known that the uncertainty computed by the retrieval (in variable *xco2_uncertainty*) underestimates the true level of error in the retrieved XCO_2, an additional term is added onto this ``theoretical'' uncertainty, in quadrature, to obtain a more realistic uncertainty per scene: σ_{SD}, the

standard deviation of all the XCO_2 values used in the 10-second average. In this *ad hoc* approach, scenes that have a very small spread in XCO_2 values across the 10-second span are assigned the theoretical uncertainty from the retrieval, while those for which the actual variability of the XCO_2 values is larger than the theoretical values are assigned a value closer to this computed error level. Both of these uncertainties are then passed through equation (39) from Baker et al., 2021 to account for error correlations. Finally, an additional error term, $\sigma_{transport}$, is added in quadrature to account for transport model errors. With all three of these terms considered, the square of the uncertainty on the 10-second XCO_2 average is given as:

$$\sigma_{10s}^2 = \frac{1}{\sum \sigma_j^{-2}} [(1-c) + c \frac{(\sum \sigma_j^{-1})^2}{\sum \sigma_j^{-2}}] + \sigma_{SD}^2 [\frac{(1-c)}{N} + c] + \sigma_{transport}^2$$

$$where\ \sigma_{SD}^2 = [N \sum X_{CO_{2j}}^2 - (\sum X_{CO_{2j}})^2] / N / (N-1),$$

and where XCO_{2j} and σ_j are the individual XCO_2 values going into the average and their retrieval uncertainties, and N is the number of good 10-second XCO_2 values in the 10-second average."

4) Line 160. "An additional term is added in quadrature to account for transport model errors. This model error term is computed from the difference between the CO2 concentrations computed by the TM5 and GEOS-Chem models when both are driven by the same realistic surface CO2 fluxes, after the annual mean difference field is subtracted off". The author referenced Basu et al., (2018) when they described the transport model error. In Basu et al., (2018), the model part error was calculated by considering the difference between a suite of inverse models optimized with in situ data and OCO-2 retrievals? Did the author do something similar here? Please clarify.

As a remark, the reference Basu et al., 2018 does not refer to the model error term but to the fact that the uncertainties between different 10s averages are assumed to be independent when assimilated. Here is the full text : *"Finally, an additional term is added in quadrature to account for transport model errors. This model error term is computed from the difference between the CO_2 concentrations computed by the TM5 and GEOS-Chem models when both are driven by the same realistic surface CO_2 fluxes, after the annual mean difference field is subtracted off; the values that result are considerably smaller than those model errors added on for the OCO-2 v7 MIP (Crowell et al. 2019). In contrast to this level of detail, the errors between different 10s averages are assumed to be independent when assimilated into the inversions (Worden et al., 2017; Crowell et al., 2019). Several studies have used this method in order to be coherent with the resolution of their inversions or simulations regarding the OCO-2 resolution (Basu et al., 2018; Chevallier et al., 2019)."*

In order to prevent confusion regarding which of the two separate points the reference refers to, we have changed the last sentence to: "Several studies have used this assumption, deeming it appropriate for the resolution of their inversions or simulations (Basu et al., 2018; Chevallier et al., 2019)".

However, to give a specific answer to the question, the model error that we use here is substantially different from that calculated in Basu et al. (2019). In that reference, the difference between the modeled CO_2 fields reflect the full impact of model differences on the inversion process as a whole, instead of just on a forward model run, since each model started by inverting the same set of in situ data. Other factors, such as the sparsity and distribution of the data, and the data uncertainties assumed, could contribute to the spread in the model CO_2 fields they obtained. In contrast, here we isolate pure transport model differences by running an identical set of CO_2 fluxes forward through two different models which are TM5 and GEOS-Chem: the difference that result are due only to differences in the models' transport (and other model features, such as resolution). The fluxes used were realistic insofar as they were taken from an assimilation of in situ CO_2 data (from CarbonTracker, 2017, Peters et al., 2007, with updates documented at http://carbontracker.noaa.gov). The resulting CO_2 mixing ratios were sampled at the early afternoon OCO-2 overflight times and a multi-year seasonally-varying climatology was developed. The deviations from this climatology from year to year were examined; and the time-varying standard deviations obtained were then used as model errors in the inversions.

5) Line 240. For LNLGv9 inversions, the available 10s OCO-2 retrievals were averaged and compared to TCCON observations". I think the author means prior and posterior XCO_2 simulated retrievals were averaged and compared to TCCON?

Indeed. We modified this sentence by adding line 256: "For LNLGv9 inversions, the available 10s prior and posterior XCO_2 simulated retrievals were averaged and compared to TCCON observations […]."

6) Line 257-258. Please indicates that Figure 5 represent only the median annual flux and the monthly median flux. In Crowell et al., 2019, the authors discussed the annual and monthly mean flux, not the median. Could you indicate why you decided to compare the median and not the mean of the flux?

The use of median annual flux and monthly median fluxes has been added in the sentence : "Figure 6 represents the median annual emissions (in PgC/yr, for the left panels) and the monthly median emissions (in PgC/month for the right panels) at the global scale for each experiment." We decided to use the median flux instead of the mean flux in order to visualize fluxes values given by most of the models. Some models might be biased for some regions for instance. We were particularly interested to not consider these outliers models and therefore to consider only the median fluxes.

7) Line 261-262: "However, the peak sinks during the Northern Hemisphere growing season (from May through September) are slightly larger with OCO-2 v7 than with OCO-2 v9". Something is missing here. In this section, you discuss global flux estimates, and at the same time, the author discusses fluxes in the Northern hemisphere. I cannot see the Northern Hemisphere Figure for this section.

In this figure, we are indeed representing the CO_2 fluxes at global scale. But the terrestrial land sink during the Northern Hemisphere growing season (from May through September) has an impact at global scale that we can observe on this figure (global scale). Impact on the seasonal cycle that have been also observed in Crowell et al., 2019 with the v7 MIP. Other previous studies also mentioned the sink during the growing season for Northern land regions (Baker et al., 2006). Byrne et al., (2017) also observed that the GOSAT nadir land data has a large sink at global scale in June-July-August due to the strong biospheric uptake of CO_2 during the Northern Hemisphere growing season. Finally, the recent study of the Global Carbon Budget 2020 by Friedlingstein et al., (2020), mentioned that the terrestrial land sink during May-Jun-July-August could be due to the combined effects of two reasons : the rise of plant growth fertilizer that increase the imput of CO_2 ; and the lengthening of the growing season in the Northern temperate and boreal areas.

In the case that other readers do not know the impact of Northern Hemisphere growing season at global scale, we've added line 297 "The large sink observed at global scale during this season is due to the strong biospheric uptake of CO2 in the temperate and boreal forest of Northern Hemisphere (Friedlingstein et al., 2020)".

8) The author includes LoFI inversion in the paper. However, there are no comments about their results. For example, LoFI seems to agree with the prior median estimate over the Northern Extra Tropics but not with IS and LNLGv9 assimilated flux.

LoFI was part of the v9 MIP not as a standard inversion but as an additional metric of flux inversion. Our study was not focusing at evaluating LoFI and comparing it with the other inversions. The goal of our paper was to compare the v7 MIP with the v9 MIP results. Consequently, we added additional information on LoFI line 82 page 4, to better explain its difference and why it was initially used in the v9 MIP : "The LoFI submission (Weir et al., 2021), new in the v9 MIP, is intended as an additional metric of flux inversion skill. LoFI uses in situ observations to match only the global atmospheric growth rate with an empirically derived land sink (Chevallier et al., 2009). The inferred fluxes are thus independent of the spatial and sub-annual variability in atmospheric observations and rely minimally, if at all, on model atmospheric transport representation. Despite the weak data constraint, it is included below with the IS inversions because it depends on the annual, global growth rate determined from observations. Given the problems flux inversions have facing remote-sensing retrieval biases (O'Dell et al., 2018) and atmospheric transport errors (Schuh et al., 2019), LoFI serves as a first-order check on inversion skill. Times and places where a flux inversion outperforms or equals LoFI's skill suggest a nominally operating system, while significantly degraded skill suggest a problem, e.g., in the prior, atmospheric transport, and/or ingested data."

9) Line 310- 313. "In the Southern Extra-Tropics, the authors mention that the sink flux estimate with OCO-2v9 is stronger than estimates made with IS, and v7, mainly because the ensemble spread with v9 is larger than with v7." I am a bit surprised that fluxes assimilated with v9 have a larger spread than v7.
OCO-2 v9 data supposed to have lower biases than v7. It seems that the error in the transport model might be the consequence of the large spread in the inversion with v9.
As a remark, and to include exactly what we have wrote in our paper, here the paragraph :
" *They show, for the whole period, stronger sinks with v9 than with IS, and v7. However, in contrast to NHExt, the ensemble spread is larger with v9 than with v7. The bias reduction of v9 gives a smaller spread and hence a better agreement among the models, particularly over the Northern Hemisphere.*"
We do not justify the stronger sink in the Southern Hemisphere because there is a larger spread among the models with v9 than with v7 in this latitudinal band. However, we do mention that the ensemble spread among the model seems to be larger over the Southern Hemisphere with v9 than with v7.
As mentioned previously for the transport model error, the error in the transport model is calculated using the differences between TM5 and GEOS-Chem and added in the uncertainties. As mentioned, the difference among the models is mainly due to differences in the models' transport. Furthermore, and compare to the v7 MIP, there are more common elements across the models in the v9 MIP where all modelers used the same *in situ* data as well as the same errors on the OCO-2 10s averages.
Over the Southern Hemisphere, there is less land cover than in the Northern Hemisphere, so we known that there are few land retrievals to constrain the land fluxes. The ensemble spread could reflect the significant uncertainty on land fluxes in this region. It could also reflect the bias in satellite retrievals due to large solar zenith angles at this latitudinal band, which indicate that there is not enough sunlight available to retrieve data. The spread between models is data-driven and reflects the models' abilities to simulate observations. All these elements could explain the lack of concordance between the models, observed mainly for this region with v9.

10) Line 314-318. Over the Tropics, the authors mention that posterior flux estimate (LNLG) is quite different to the prior estimate (Fig 6.c), but not comments are provided. How reliable can be the posterior estimate over the Tropics knowing that OCO-2 retrievals can be biased due to cloud coverage during the wet season and aerosol from biomass burning during the dry season, as the author mentioned?
The lack of validation data over the tropics make this comment difficult to answer. As we mentioned in the paper, there is more OCO-2 data in tropics than the IS data but the OCO-2 data are biased during the dry season due to aerosol from biomass burning and due to cloud during the wet season. There is not enough validation data as well to estimate which from the OCO-2 or the IS posterior emissions are

more reliable. Some aircraft project effort organized by the National Oceanic and Atmospheric Administration are ongoing to address these biases.

11) Could you explain a bit more about the dipole in northern Africa? You mentioned several studies, but a better explanation is needed.

In order to give more details about this dipole between Europe and Northern Africa we changed our sentence page 20 line 366 :

"Inferred fluxes using v9 seem then to be more consistent with other studies, but more analysis is needed to understand why this difference between v7 and v9 appears over Europe (which could be due to a dipole between Europe and Northern Africa as observed and mentioned by the previous studies of Houweling et al. (2015), Chevallier et al. (2014), Feng et al. (2016), Reuter et al. (2014), and Reuter et al. (2017)) but not for Northern America and Northern Asia. Previous studies already observed and mentioned a larger European land sink in balance with a large tropical land source. Particularly, Houweling et al. (2015) found a difference in flux between these two regions of around 0.8 PgC/yr. They found that this balance was caused by a lack of GOSAT observations during the winter over Europe. Additionally, Chevallier et al. (2014) also observed this balance between Europe and Northern tropical Africa in their GOSAT inversions and they considered the large source over North Africa has unrealistic. According to Feng et al. (2016) the large sink over Europe inferred from GOSAT data was caused by large biases outside of the region, which for mass balance, the inversions was removing larger CO2 over Europe, in agreement with Reuter et al. (2014) and Reuter et al. (2017)."

12) Fig 8 (NH Tropics) and (SH Tropics) LoFi monthly seasonally seem to be offset? Could the author explain why LoFi might not be capturing the seasonally over these latitudes bands?

As mentioned above with more detail for the comment (9), LoFI uses a different method than the other inversions but fit some independent data as the other simulation do. LoFI has then been used in this MIP project to look at a range of different methods.

LoFI is not considered as a standard inversion compared to the other models, it is a bit like an outlier. And the comparisons to *in situ* data (Figure 11) show that there's probably some degraded skill for LoFI in the tropics, but not much due to the significant uncertainties there. In the Fig. 3 of Weir et al., 2021, LoFI is compared to inversions with a broader collection of priors than used in the MIP. It does have a bit of a phase shift compared to these inversions but it is within the uncertainty range when compared to all the TRENDY v7 models.

13) Line 412. Alternatively, the OCO-2 inversions could be dominated by savanna seasonality (Baker et al., 2021a in prep). I cannot find (Baker et al., 2021a) in the reference list, and the reference cited (Baker et al., 2021) does not mention anything about the savanna seasonality but how to calculate the error correlations in OCO-2 data. Please provide the correct reference.

This referred to a paper in preparation by Ian Baker. The paper has not been submitted yet. We removed the year in the citation and added the first name of the author for no confusion with the paper of David Baker et al., 2021.

14) Line 426-. "Over the southern hemisphere, a large underestimation of the ensemble mean of LNLGv9 appears compared to the observations". I am surprised with these findings; I would have thought that the spatial coverage that OCO-2 data in the southern hemisphere would improve the results compared to IS. I am also surprised by the significant difference between LNLGv9 and IS biases in SH. Do you know how large are LNLGv9 OCO-2 biases in SH compared to v7? You mentioned in the introduction that biases in LNLG v9 were reduced considerably compared to v8 and v7.

Figure 12 shows the normalized bias and standard deviation for the ensemble mean of IS and LNLGv9. For the biases evaluation with the withheld data, we can see an increase in the biases going from the Northern latitudes (of maximum 0.5) to the Southern ones (reaching -3.5 for the LNLGv9). The variability seems to follow as well, where the variability is larger over the Southern Hemisphere than the Northern Hemisphere. This shift of variability and biases between the Northern and Southern Hemispheres seems to be linked with the number of withheld data. We can see Figure 2 that the number of withheld data is lower over the Southern Hemisphere with less than 1000 data, while the number of observations is above 10 000 for the Northern Hemisphere. This large underestimation observed with LNLGv9 over Southern Hemisphere when compared to the withheld data could be explain with the lack of withheld data over this Hemisphere.

We added this sentence page 27 line 461 : "In addition, this same variability seems to be disproportional to the number of withheld data (see Fig. 2). Indeed, the standard deviation for both IS and LNLG is low when the number of withheld data is important (superior to 10 000 data)."

Furthermore, the MDM values and withheld data were not considered during the v7 MIP. This attempt to account for correlations between the MDM errors was made only for the v9 MIP. Evaluating v7 with the withheld data can not be performed. However, the biases in LNLGv9 have been reduced considerably compared with v8 and v7 when evaluated with the TCCON data as discussed in the TCCON section (3.4.3).

15) Is it possible to provide a Table in the appendix with no normalized bias?

We thought more relevant to use normalized values for this evaluation as the ranges of data among the models can be large and different. Particularly, since the MDM values range over two orders of magnitude, the use of the normalized residuals gives the most meaningful interpretation of the residuals. We decided then to keep the normalized values for better comparison of scale variability. Additionally, more information on the scaling can be found with the RMSE (root-mean-square error) values in ppm. Using RMSE we can estimate how much the models differ from the withheld data.

16) Line 466. It is possible then that this excess of concentration, in both experiments, reflects the initial conditions of the inversion. Is there any way to test this?

One possible way to verify this would be to have vertical profile of convective mass fluxes comparing the meteorological conditions first (such as the convection from ERA-Int, GEOS-FP or MERRA-2 for instance). Schuh et al., 2019 suggested that the differences and biases observed between TM5 and GEOS- Chem could result from the representation of vertical motion. Ongoing studies are focusing on this difference, such as Schuh et al., (in prep) where their new finding suggest difference mainly coming from the meteorological input.

17) Fig 13 and Fig 14. I don't understand why the author normalized the biases. It is clear that RMSE is large at some latitudinal bands and TCCON sites, suggesting that raw biases might also be high. I would also consider adding this information to the manuscript.

We are not sure to really understand what the reviewer means for this comment regarding the TCCON sites. Figure. 13 and Fig. 14 are showing normalized biases for the withheld data and we justified in the previous comment (16) and in the main text (line 470 page 28) why we are using the normalized biases when evaluating with the withheld data.

18) Line 471-474. "Compared to the evaluation of ISv7 in the study of Crowell et al. (2019), ISv9 and LNLGv9 biases are closer to each other (in the v7 MIP, the LNv7 were biased high compared to ISv7). Additionally, the OCO-2 biases have decreased (to values between -1.0 and 1.0 ppm) with v9 compared to v7, where biases ranged between -1.5 and 1.5 ppm (Crowell et al., 2019)". I don't understand why the authors say that OCO-2 biases have decreased compared to Crowell et al., 2019. Did Crowell et al., 2019 normalize the bias? If not, I don't quite understand the comparison.
As comment, the biases are normalized only for the withheld evaluation, they are not normalized for the TCCON evaluation. This paragraph refers to the TCCON evaluation and to figure 17 where the biases are not normalized. We then were able to compare our TCCON evaluation with the one used in Crowell et al., 2019.

19) Line 474-476. "…to the accuracy of TCCON retrievals over these regions (Crowell et al., 2019). I think that biases might likely be associated with biases to satellite observations than TCCON bias. Besides, I dont think that Crowell et al., 2019 is not a good reference for talking about the accuracy of TCCON. XCO2 TCCON retrievals can contain airmass-dependent biases, which are corrected using the method described by Wunch et at.. Just wondering how wrong could be this correction to cause the posterior concentration biases seen here.
Even if this has been mentioned in Crowell et al., 2019, we removed this comment in order to give two assumptions for this positive bias observed over most of the European sites. We changed consequently line 504 page 30 : "As observed here as well, IS and v9 have large positive biases over most of the European sites, which could indicate either an issue related to the coarse resolution used by the transport models or to a latitudinal bias (though this is not shown here, positive biases are also observed for the East Trout Lake TCCON site situated in Canada at almost the same latitudinal band as the European sites)".

20) Line 476. Could the authors provide the location of the TCCON sites in a map instead of Table 3? It would be better for the reader to see where Caltech, Saga and Tenerife TCCON sites are located. I had to look at their locations from other manuscript.
We provided a map of the TCCON sites location and included it page 13 with the sentence line 258 page 13 "All TCCON sites used in the evaluation section are listed in Table 3 and Fig. 3 represents the location of the TCCON sites".

21) The Caltech, Saga, and Tenerife sites show large underestimations in the IS and v9 results across all models. Isn't it bad to find a large underestimation at Caltech for the reliability of the fluxes estimated in North America? Is the location of Caltech a coastal site in the model? If so, it might strongly affect by ocean fluxes where not OCO-2 observations were assimilated.
As mentioned in our paper, Caltech shows indeed large underestimation in the IS and v9 results across all models, however Edwards site which is very close to Caltech shows lower bias with either underestimation or overestimation according to the models used. Caltech is not the only one site representative of the Northern America however. Lamont situated in Oklahoma and Park Fall in Wisconsin, are also part of the Northern America TCCON sites.
Regarding the comment that Caltech might be affected by ocean fluxes where no OCO-2 data are assimilated, this might not be a reason for the underestimation observed with IS and LNLG simulations. Indeed, when we look other coastal TCCON sites like Rikubetsu, Tsukuba or Ascension Island, they do not show large underestimation but a slightly overestimation. We discuss about these other sites in the paper. Besides, as we already mentioned in the paper and observed in v7 MIP as well,

differences between the Caltech and Edwards sites (which are very close each other) could be due to the location of Edwards over the mountains while Caltech is affected by the Los Angeles basin (Kort et al., 2012; Schwandner et al., 2017). So, the coarse resolution of models cannot differentiate the variability of these two sites (Crowell et al., 2019; Schuh et al., 2021).

22) Line 480. Another possible explanation of the underestimation observed over Saga and Izaga is that these small islands are strongly influenced by ocean fluxes, where the assumed uncertainties are small compared to land.

We thanks the reviewer for this comment. This underestimation could indeed be linked to the fact that they are small islands. But the bias observed in Izana could also be linked to the high altitude of the TCCON site. We modified the paragraph, line 480 page 27, by :

"This could also explain the underestimation observed over Saga and Paris, which are urban regions. However, Saga is also a small island and could hence be influenced by ocean fluxes, where the assumed uncertainties are small compared to land. The underestimation observed for Izana (Tenerife Island) is probably linked to the same uncertainty (being a small island) but could also be due to the high altitude of the site."

The biases in v9 have decreased for Ascension Island compared to v7, where the biases were around 1.0 ppm for LN and LG. Are these results compared to TCCON biases presented in Crowell 2019? If so, I don't understand why the authors compare standardized bias against not standardized biases?

In our v9 MIP evaluation with the TCCON data, we presented bias and standard deviation for all TCCON sites by model. All evaluations in our paper are only for the v9 simulations. We did not include v7 simulations as the evaluation of the v7 MIP is already presented in the paper of Crowell et al., 2019. We do not understand, why the reviewer mentioned that a comparison between standardized bias against not standardized biases has been done in the paper when it is not the case. The way we did our evaluation with TCCON is similar to what have been done in the paper of Crowell et al., 2019.
As a remind also, biases were only normalized for the withheld evaluation.

23) Line 495. "Transport model uncertainty is not expected to have changed dramatically since v7. This suggests that the reduction in the ensemble spread is likely related to a decrease in OCO-2 retrievals errors in v9 compared to v7". Are the modellers that participate in OCO-2 MIP have trying to consider improvement in the transport modelling (at the surface)?

We thanks the reviewer for this comment. All of modelers did not modified or updated their transport model between the two MIP versions, except the simulation CAMS. The transport model used in CAMS is different between the two versions. The meteorology has been updated with ERA5 instead of ERA-Int. Then, the vertical mixing is different between the two version where it has been changed from Tiedtke., (1989) to Emanuel., (1991) convection scheme with addition of thermals terms. But further work is needed to demonstrate which of transport version performs the best.
Modelers participating in the OCO-2 MIP are involved in efforts to discern the *causes* for differences in the two most common CTMs used for the MIP, TM5 and GEOS-Chem. Experiments, and corresponding papers (Schuh et al., in prep) are underway focusing on differences in parameterized vertical transport. This work is still at the point of discerning specific causes for differences, as opposed to fixing any known issues with either model. Therefore, we are left with the assumption, for the time being, that the differences between the CTMs, including near surface behavior as pointed out, arise primarily as differences in the parent model meteorology that drive both models.

24) Line 525-526. "When they compared..." who is they? I think the reference is missed here.

"They" was referring to the study of Gloor et al., 2018. We added this reference in the sentence.

In the discussion section, I think it is also important that the author indicates why the assimilating OCO2 LNLGv9 data over the Tropics shows a stronger seasonality than prior fluxes. For the reader, it would be good to know why prior fluxes derived from ORCHIDDE or CARDAMON are not capturing well the seasonality. What could be wrong with these models?

Figure 9.b and 9.d shows the monthly median fluxes for the different simulations and prior over Northern Tropics and Southern Tropics respectively. We can see on this figure that the prior (in black) has the same seasonality than the OCO-2 LNLGv9 (in dark blue) but a difference in magnitude. The sources are particularly lower with the priors than with the LNLGv9. For information regarding the priors used, only two models used ORCHIDDE or CARDAMON, the rest used CASA-GFED.

Net fluxes in the tropics are sensitive to the Gross Primary Productivity (GPP) and the Respiration (R). However, large fluxes are present in the tropics which can give large GPP and R in this region. Consequently, the way they are calculated in these models can bring not negligible differences between the priors and the posteriors. Additionally, these two lands models are parameterized mainly with Northern Hemispheric information and so might not represents correctly the tropical latitudes (Ian Baker et al., in preparation). More information on the ORCHIDDE, CARDAMON and CASA-GFED comparison will be provided in this future paper, which was not the subject of our paper.

25) Line 552 -554 "…with a slight negative bias in the v9 OCO-2 data for almost all latitudes, particularly in the Southern Hemisphere and the tropics, where few evaluation data are available". Here the author mentions "slight negative bias in the v9 OCO-2 data", however in lines 426-427, they write, "Over the southern hemisphere, a large underestimation of the ensemble mean of LNLGv9 appears compared to the observations". This seems a bit contradictory.

The results, in the evaluation section using the withheld data, show large negative bias in the Southern Hemisphere but slightly negative over the tropics. We modified line 552-554 with :

"Evaluation with the withheld data, ATom aircraft measurements, and TCCON retrievals suggests similarities in biases between the in situ data and LNLGv9 data, with negative bias in the v9 OCO-2 data for almost all latitudes, particularly large in the Southern Hemisphere and slightly negative in the tropics, where few evaluation data are available. Evaluation against TCCON shows also a reduction in retrieval errors with v9 ensemble models compared to v7."

Editorial and minor comments:

Title: Four years of global carbon cycle observed from OCO-2 version 9 and in situ data, and comparison to OCO-2 v7. As a personal opinion, I would write version 7 instead of v7 to be consistent with "version 9".

We considered this remark and change the title.

Line 121. "Inversions assimilating OCO-2 ocean retrievals produced unrealistic results with annual global ocean sinks higher of 2.6 ± 0.5 GtC.yr−1 compared to the state-of-the-art estimated in Le Quéré et al. (2018)." Could you quote the ocean sink estimated in Le Quéré et al. (2018).

The ocean sink estimated in Le Quéré et al. (2018) has been added in the sentence : "inversions assimilating OCO-2 ocean retrievals produced unrealistic results with annual global ocean sinks higher of 2.6 ± 0.5 GtC/yr compared to the state-of-the-art estimated in Le Quéré et al. (2018), which was of 2.5 ± 0.5 GtC/yr in 2017."

Line 136. (Kiel et al. (2019)). Remove parenthesis to 2019. We removed the parenthesis.

Line 151. "Details of the form and derivation of these average uncertainties may be found in the 'constant correlation' section of Baker et al. (2021).", Please be consistent with the quotation mark "constant correlation". We changed the quotation mark.

Line 137. "This bias correction in v9 allows a more uniformly. Replace the word uniformly by uniform. We replaced it.

Line 184. There is an extra point. Please remove. Extra point removed.

Line 241-242 Figure3. Please add the y-axis description (right legend). The y-axis description has been added.

Line 311. "However, in contrast to NHExt". In this line, you need to define define NHEx.
We added it : "However, in contrast to the Northern Extra-tropics (NHExt)"

Line 305: The Southern extra-tropics. Capital letter?. We changed it as well as everywhere else in the text.

Line 306. Correct the word "signifcaintly" to significantly. We modified it.

Line 398. For the monthly emissions of the Southern Tropical regions (Fig. 10), we can see the strong impact of El Niño in fall 2015 over Southern Tropical Asia in the larger emissions (with a maximum of 0.35 ±0.01 PgC/yr) given by all the inversions compared to the rest of the period. Please re-write, it is hard to understand.
We modified this sentence by :
"Looking at the monthly emissions of the Southern Tropical regions (Fig. 10), we can see the strong impact of El Niño between August and November 2015 over Southern Tropical Asia. The emissions reach a maximum of 0.35 ±0.01 PgC/yr, highest of around 0.30 PgC/yr compared to the rest of the period."

Line 400. This large fall 2015 mainly come from Indonesian fires. I would write: The large emissions from Southern Tropical Asia (Fig.10.f) primarily come from Indonesia fires. I would also remove the cyan and green lines from all plots (Fig.5 to Fig.10) that show monthly median fluxes for 2018-2019. If there is no data there, I don't think it is needed to show that the median is zero.
We considered this suggestion and modified the sentence accordingly. We also changed all figures in order to remove the median values when they are equal to zero.

Line 605. … is about 0.5 GtC.yr−1.Land (need space) We added a space.

References :

Baker, D. F., et al. (2006), TransCom 3 inversion intercomparison: Impact of transport model errors on the interannual variability of regional CO_2 fluxes, 1988–2003, Global Biogeochem. Cycles, 20, GB1002, doi:10.1029/2004GB002439.

Byrne, B., D. B. A.Jones, K. Strong, Z.-C. Zeng, F. Deng, and J. Liu (2017), Sensitivity of CO_2 surface flux constraints to observational coverage, J. Geophys. Res. Atmos., 122, 6672–6694, doi:10.1002/2016JD026164. Received

Friedlingstein, P., et al. (2020), Global Carbon Budget 2020, Earth System Science Data. Issue 12, Pages 3269-3340, doi.org/10.5194/essd-12-3269-2020.

Tiedtke, M., 1989: A comprehensive mass flux scheme for cumulus parameterization in large-scale models.*Mon.Wea. Rev.*, **117**, 1779–1800, doi:10.1175/1520-0493(1989)117<1779:ACMFSF>2.0.CO;2.

Emanuel, K. A., 1991: A scheme for representing cumulus convection in large-scale models. *J. Atmos. Sci.*, **48**, 2313–2329, doi:10.1175/1520-0469(1991)048<2313:ASFRCC>2.0.CO;2.