Response to Reviewer 1

Reviewer's comment is in black. Our responses are in blue text. The references are to the manuscript with changes tracked.

General comments: the authors have developed an integrated greenhouse gas transport system from the global to regional scales. In this study, the regional domain of interest they chose is Canada and the United States. Three experiments were set up and evaluated with various meteorological and CO2 data. These experiments are GLB90, GLB45, LAM. GLB90 is the global model output with 0.9 deg grid spacing, GLB45 is similar to GLB90 but with 0.45 deg grid spacing, and LAM is the regional setup at the 10 km resolution driven with GLB45' meteorology and CO2. All of the experiments used the same surface fluxes from CT2016. That the main message of this study delivers is the high-resolution outperforms the simulations. The content of the manuscript lies in the category of the model description papers of GMD and meets the requirement of this category. The manuscript is generally well written, and the analyses presented are fairly well, I would like to recommend for publication after addressing my concerns as follows.

Response: We appreciate the reviewer's careful reading of our manuscript and helpful comments. We have revised the manuscript following the reviewer's suggestions.

1. The CT2016 posterior CO2 mixing ratios are used as a baseline in Figure 7. It seems to me that the transport difference causes much bigger discrepancy than those by grid spacing. Both of Carbon Tracker and GEM-MACH-GHG are an operational system, and this is a forefront work towards (regional) flux estimates. One recent study on the transport uncertainty (Schuh et al, 2019) demonstrates the large difference in CO2 mixing ratios between TM5 and Geos-Chem models using the same CT surface fluxes. They show that the inverse model flux estimates for large zonal bands can have systematic biases of up to 1.7 PgC/year due to large scale transport uncertainty. The difference between CT2016 and GEM-MACH-GHG models in Figure 7 implies the bias in the GEM-MACH-GHG transport. I would expect the authors to have a discussion on the possible strategies to improve GEM-MACH-GHG transport since it is an operational system as Carbon Tracker. Additionally, my personal experience of working with CT2016 and CT2017 is that CT2017 outperforms the posterior CO2 mixing ratios a lot over CT2016. I encourage the authors to add CT2017 results in Figure 7 as well.

Response:
There are actually three separate points being made by the Reviewer in this first comment. We will address all 3 points in the next 3 paragraphs.

One point concerns the mismatch of GEM-MACH-GHG and TM5 transport errors. While the Reviewer states that the difference between 2 models "implies a bias in the GEM-MACH-GHG transport", in fact what we can infer is only a mismatch in transport between the two models. The reason for the mismatch was already discussed in Section. Page 17, lines 26-29: "As discussed in Polavarapu et al. (2016), GEM has different transport behaviour from the transport model used in CarbonTracker, in particular over the Arctic region, as seen in time series of CO_2 concentrations and column-averaged CO_2. Thus, our models are not expected to perform better than CT2016 because we use surface CO_2 fluxes inferred by an inversion framework using a different transport model which has different transport behaviour". All experiments

(GLB90, GLB45 and LAM) used posterior CO_2 fluxes from CT2016 that are inherently consistent with the transport behaviour of TM5 which is the transport model used in CT2016. Understandably, posterior CO_2 mixing ratios in CT2016 take advantage of this consistency, producing better results (smaller biases in general) than the GEM-MACH-GHG. This baseline result is included just to show that our model behaviour does not diverge too much from the CT2016 result despite this handicap. Of course, the main purpose of this study is to report the performance of the developed regional version of GEM-MACH-GHG compared to its (global) variations. Our strategy to properly assess the transport behaviour of the GEM-MACH-GHG in estimated fluxes is to develop a flux estimation system based on the GEM-MACH-GHG as mentioned in the Section 5 (Page 18, lines 2-4). Only then will our model simulations with posterior fluxes be consistent with the model dynamics used to generate the posterior fluxes, and the comparison to observations fairer. This development is currently in progress.

Another point concerns the use of CT2017 instead of CT2016 in Figure 7. The correct comparison for our Figure 7 is CT2016 because our model actually used CT2016 fluxes to produce the mixing ratios. Nevertheless, following the reviewer's suggestion, we looked at CT2017 results in addition to CT2016 results in the same style as Figure 7 in the manuscript (Figure R1), excluding the GEM-MACH-GHG results to focus on the comparison between two CT versions. It is difficult to say that CT2017 outperforms than CT2016 with respect to the posterior CO_2 mixing ratio, in terms of monthly bias and its standard deviation, at least for the afternoon (12-16 LST) data in 2015. On the other hand, looking at the results for every hour data (Figure R2), CT2017 outperforms CT2016 in terms of CO_2 bias, as the Reviewer expected. However, since we used fluxes from CT2016 which assimilated afternoon data in its inversion, the appropriate comparison is to CT2016 and afternoon time data in Figure 7. Since it would further crowd Figure 7 and distract from the main point of the figure, we decided not to add CT2017 results in figure 7. Instead, we have revised the manuscript to better clarify the motivation of using CT2016.

Page 11, lines 20-24: The CO_2 fields in the LAM and other experiments are investigated in terms of monthly bias and STDE of daily afternoon (12:00-16:00 LST) modelled CO_2 concentrations at the measurement sites shown in Fig. 1 and listed in Table 1 (Fig. 7). Daily afternoon time was selected because this is what CT2016 used to estimate surface CO_2 fluxes. Also, CT2016 results are included as a reference ~~in order to verify results of three experiment simultaneously~~since this is what all our model experiments used as input fluxes.

A final point concerns the fact that transport error may be large compared to error of model resolution. On this point, we concur. In the future, our flux estimation system development will solve the problem of the mismatch of transport error with TM5. However, we will still have the problem of transport error since every transport model has this issue. In the future, we plan to address this issue by using different sources of meteorology in the flux estimation system. This point is now mentioned in the discussion section.

Page 18, lines 3-7: When posterior fluxes become available from our global model, this will alleviate the issue of model transport error mismatches with CarbonTracker~~being embedded in the surface fluxes~~. However, we will still have transport error, which is one of the biggest sources of posterior uncertainties in an inversion (Schuh et al., 2019). To address this issue, we plan to use multiple sources of meteorology to better account for transport error in posterior flux and uncertainty estimates.

Schuh, A. E., Jacobson, A. R., Basu, S., Weir, B., Baker, D., Bowman, K., Chevallier, F., Crowell, S., Davis, K. J., Deng, F., Denning, S., Feng, L., Jones, D., Liu, J., and Palmer, P. I.: Quantifying the Impact of Atmospheric Transport Uncertainty on CO_2 Surface Flux Estimates, Global Biogeochem. Cy., 33, 2018GB006086, https://doi.org/10.1029/2018GB006086, 2019.

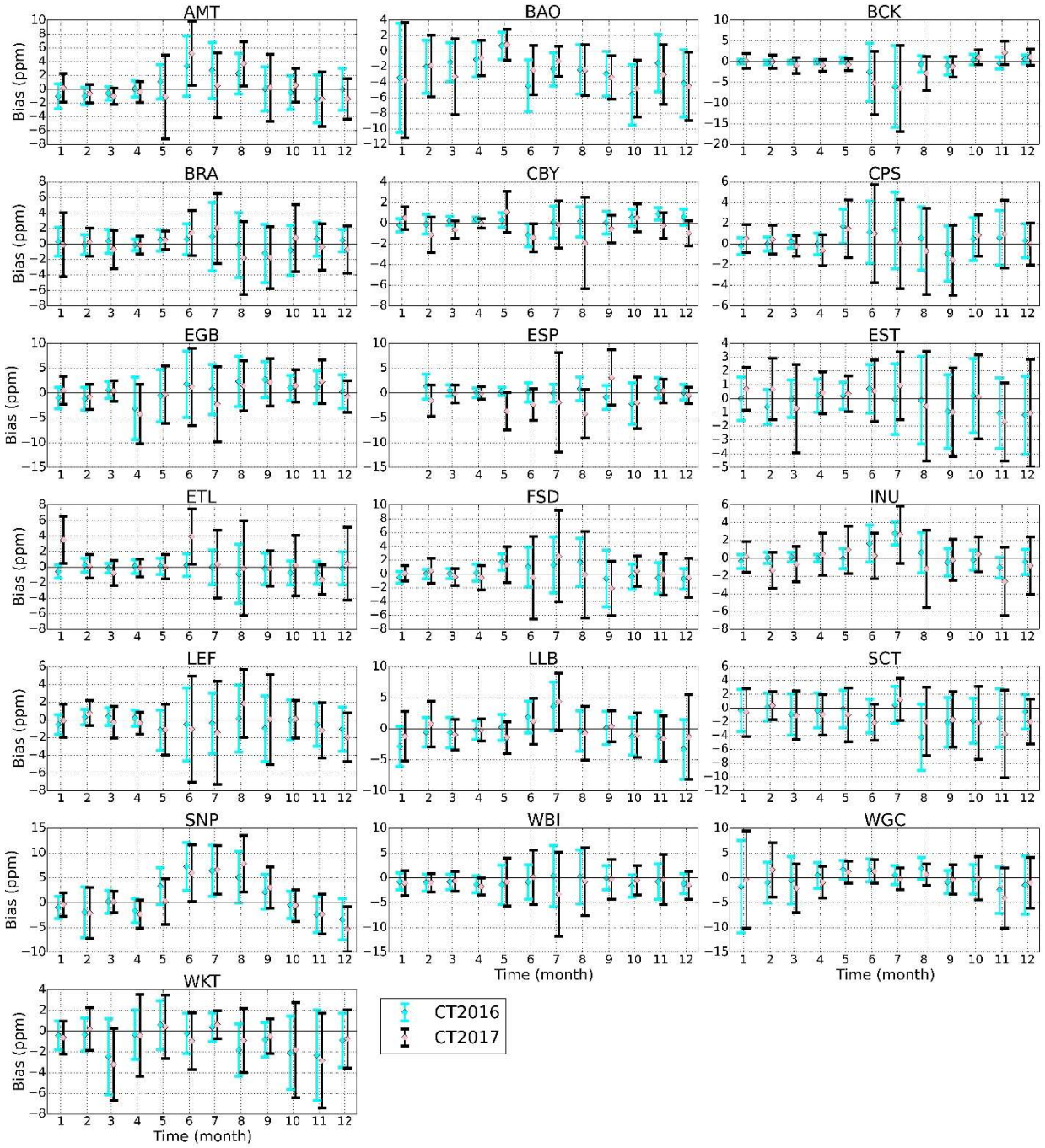

Figure R1: Monthly mean bias and STDE of daily afternoon averaged (12:00-16:00 LST) CO_2 concentrations from CT2016 (cyan) and CT2017 (pink marker with black error bar or STDE) at all measurement sites used in this study.

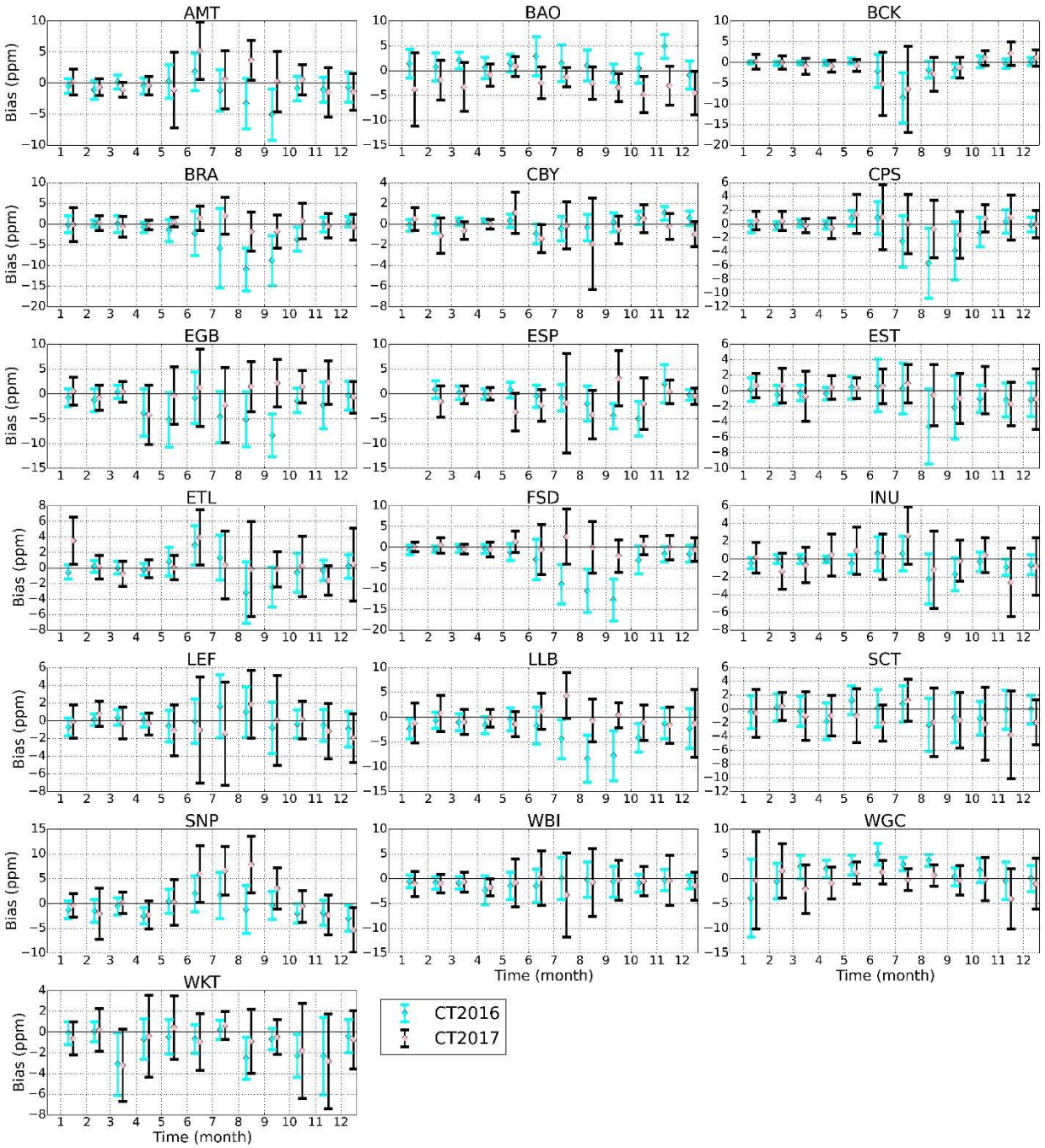

Figure R2: Same as Figure R1, but for observations minus model residuals computed every hour. Note the y-axis scales vary with location.

2. The authors use site codes quite often in writing. It may cause difficulties and confusion to some readers who are not familiar with the geographic locations of those sites to follow the statements. I recommend the authors to add the codes of the stations in Figure 1 for the references.

Response: We fully agree with the reviewer's comment. Thus, we have added site codes to Figure 1.

3. In the manuscript, the authors gave the credit to LAM for resolving the complex terrain better. This is an overstatement. I don't see clear improvement at the mountain sites (such as BAO, in Figure 9 and 10) or coastal sites (such as ESP, in Figure 9 and 10). I was puzzled by that because I also expected the outperformance from the high resolution simulations. One of the reasons that the improvement of the high-resolution simulations is not the evidence is that the analyses/figures shown were averaged over a relatively long period (monthly or seasonally). The improvement should be more evident in a higher frequent timescale, such as daily or even hourly due to resolving faster physics. I recommend the authors to select a or a few cases and to show hourly/daily time series for that/those case(s), as Figure 15 in Agustí-Panareda, et al., 2019, by which the authors should be able to justify the statement better.

Response:
To examine shorter timescales, Figure R3 shows time series of the residual of modelled CO_2 from hourly observations for all three experiments. Indeed, LAM has smaller residuals over the global GEM-MACH-GHG for certain specific periods at ESP. However, BAO residuals are still similar for the 3 models. Thus, it is possible to see the improvement of the high-resolution simulations at ESP by looking at the higher-temporal frequency data, as suggested by the Reviewer. We added Figure R3 in the supplementary section as Figure S1 to show an example of better CO_2 simulation by the regional model and revised the manuscript accordingly.

Page 13, lines 10-13: At many sites, the benefit of finer grid spacing is evident, but higher horizontal grid spacing does not always guarantee a lower magnitude of bias for all sites. Part of the reason that improvement is not clear at certain sites in Figure 9 may be due to the focus on afternoon mean values. For example, if we consider higher temporal frequency output (i.e. hourly residuals), the LAM is better than GLB90 and GLB45 even at ESP in November (Fig. S1).

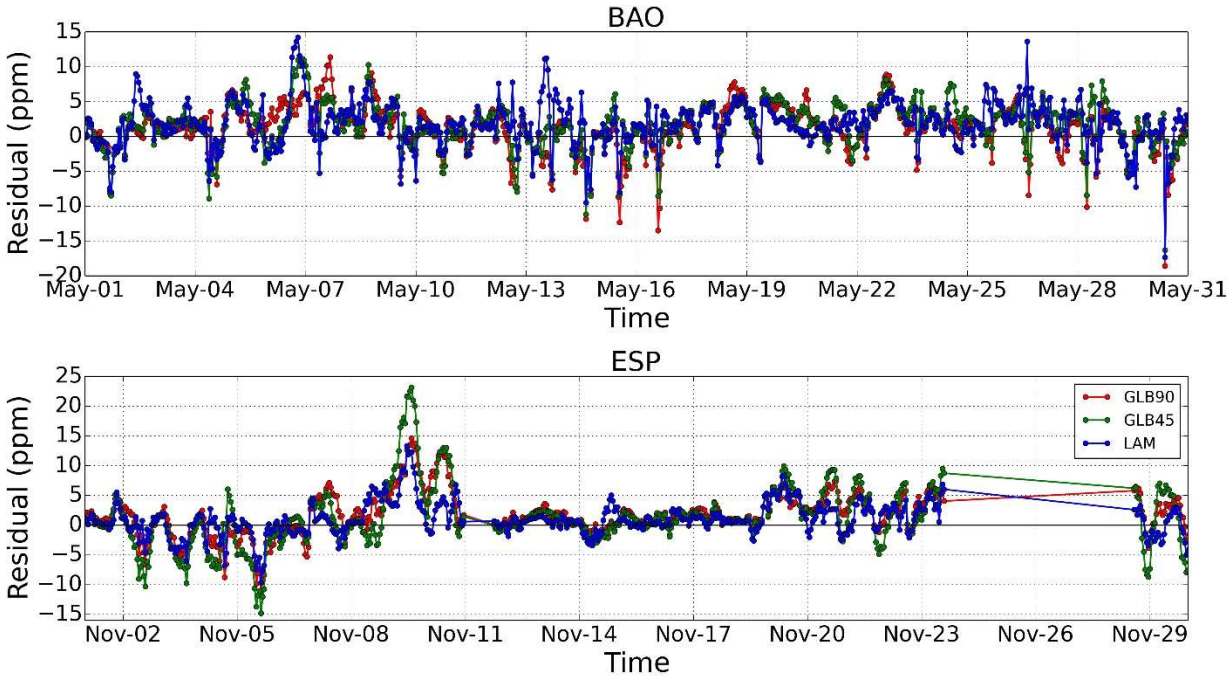

Figure R3: Residual of hourly modelled CO_2 (ppm) time series and observations at BAO (Model-Observation) from GLB90, GLB45 and LAM experiments for May 2015 (top) and ESP for November 2015 (bottom).

Specific comments:
1. P2/L12, add a more recent work from OCO2MIP (Crowell et al., 2019).

Response: Following the reviewer's comment, we have added Crowell et al. (2019) in the suggested place.

Crowell, S., Baker, D., Schuh, A., Basu, S., Jacobson, A. R., Chevallier, F., Liu, J., Deng, F., Feng, L., McKain, K., Chatterjee, A., Miller, J. B., Stephens, B. B., Eldering, A., Crisp, D., Schimel, D., Nassar, R., O'Dell, C. W., Oda, T., Sweeney, C., Palmer, P. I., and Jones, D. B. A.: The 2015–2016 carbon cycle as seen from OCO-2 and the global in situ network, Atmos. Chem. Phys., 19, 9797–9831, https://doi.org/10.5194/acp-19-9797-2019, 2019.

2. P16/L20, "from Figure 15, it can be concluded that higher horizontal resolution might help to enhance the performance of CO2 simulations even without using fluxes on a finer grid spacing" is an overstatement. I don't see the statement is true. Please clarify it. Even though the observation network is sparse, there are still a few sites in the domain of interest. I would like to see the same matrices from the observations overlaid to the model values in Figure 15.

Response:
We agree with the Reviewer that unwarranted conclusions were made from this figure. Figure 15 basically shows that the regional model can generate fine-scale spatial variations of CO_2, in particular the diurnal cycle of CO_2 over the complex terrain. This is thanks to better resolving topography in the regional model and to the presence of finer spatial scales in weather forecasts. However, though we expect such fine scale variations in the diurnal cycle amplitude to exist, we cannot say that this particular pattern is correct because there simply isn't the observation density to verify these spatial scales. For that reason, Figure 15 has been removed from the manuscript and moved to the supplementary as Figure S3, and we have revised the manuscript accordingly.
As explained in the specific comment #10 below, due to the nature of the method we used for extracting the amplitude of signals from hourly to a longer time scales, any gaps in data is not allowed. Nevertheless, we did compare not only diurnal but many other frequencies to observations at a few sites. Figure 14 shows an example of the same matrix used in the Figure 15 (now Figure S3 in the revised manuscript) for the site where observation data is available for the same period as in Figure 15. Lines at 1D on the x-axis correspond to the amplitude of diurnal cycle of modelled or observed CO_2 concentration for the period from June to August 2015. At the lowest measurement elevation the LAM is closest to observations for not only the diurnal cycle but almost all frequencies.

Page 1, lines 17-20: Better representation of model topography in the regional model ~~reduces transport and representation errors significantly~~ results in improved simulation of the CO_2 diurnal cycle compared to the global model~~, especially in regions of complex topography, as revealed by the more precise and detailed structure of the CO_2 diurnal cycle produced at observation sites and in model space~~ at Walnut Grove, California.

Page 16, lines 10-34: The amplitude of the diurnal cycle can also be computed in model space to illustrate its spatial variability as a function of model resolution (Fig. S3). With the same prescribed fluxes, greater spatial heterogeneity of diurnal cycle amplitude occurs with increased resolution. However, the validation of these finer spatial scales requires a dense observation network and is not possible at present.

~~Because the observation network used in this study is rather sparse, we also compute the amplitude of the diurnal cycle of CO_2 in model space instead of observation space to illustrate its spatial variability. Figure 15 shows the amplitude the diurnal cycle of CO_2 (i.e. 1-day period) over land regions in western and eastern North America using the same method used in Fig. 14. Hourly modelled CO_2 concentrations at the lowest model levels from the three experiments during JJA are selected. Results over oceans and northern Canada are excluded as these regions have much smaller amplitudes relative to those shown in the figure. Even though the same coarse resolution surface CO_2 fluxes (originally on a $1° \times 1°$ grid) are used in CO_2 simulations for all three experiments, the LAM experiment produces a more detailed spatial pattern than either of the GLB90 and GLB45 experiments on account of its better representation of topography and the variability of simulated PBL height (which is related to the vertical mixing of tracers within PBL, in particular over the Rocky, Sierra Nevada and Appalachia mountains, where terrain-induced circulation also plays an important role (De Wekker and Kossmann, 2015), and along coastlines). The spatial pattern of the diurnal cycle is strongly correlated with that of surface biospheric fluxes (not shown) and model topography (Fig. 1). On the other hand, the impact of fossil fuel fluxes on CO_2 diurnal cycle is small because it has a consistently positive value although prescribed fossil fuel fluxes included in optimized flux from CT2016 have day of week and diurnal variations. Figure 15 also suggests that better representation of the diurnal cycle of CO_2 concentrations in the LAM experiment extends beyond the WGC site that was shown as an example in Fig. 13. Unfortunately, due to the limited observation coverage, it is not possible to verify this pattern over the whole domain using observations, at the moment. Nevertheless, from Figure 15, it can be concluded that higher horizontal resolution might help to enhance the performance of CO_2 simulations even without using fluxes on a finer grid spacing. This is the case when comparing the CO_2 diurnal cycles between the GLB90 and GLB45 experiments. This hypothesis is consistent with the recent finding that higher horizontal resolution global models can simulate more detailed spatial patterns of the CO_2 diurnal cycle compared to low horizontal resolution global transport models (e.g. Agustí-Panareda et al., 2019).~~

Page 18, lines 17-18: ~~Significant~~Noticeable improvement in reproducing the CO_2 diurnal cycle by the regional model can be seen at ~~sites~~WGC which is located in complex terrain ~~region~~.

Page 19, lines 6-8: Since the regional model can simulate the spatial heterogeneity of the diurnal cycle of CO_2 in model space (Fig. ~~15~~S3), better observational density is needed to ~~distinguish~~validate the performance of CO_2 simulations in the regional model in more detail.

3. P17/L28, "This is a promising result because it suggests that using night-time data in an inversion to estimate night time fluxes (e.g. Lauvaux et al., 2008) may be beneficial if a high-resolution model is used." This can be demonstrated by using CT2017. CT2017 has assimilated nighttime data, but CT2016 doesn't.

Response:
Yes, as shown in Figure R2 above, CT2017 can demonstrate this. Thus, this statement is corroborated. Since CT2017 was not yet released when this model was being developed we used what was available at

that time, namely, CT2016. In the future, when assessing our flux estimation system (currently under development) for the LAM, comparisons to CT2017 are warranted. The statement the Reviewer indicated aimed to compare our regional GEM-MACH-GHG with global GEM-MACH-GHG to demonstrate the potential benefit of a higher horizontal resolution in future inversion study. To clarify this point, we have revised the manuscript.

Page 17, line 30 – page 18, line 3: ~~However, despite this handicap, beyond afternoon time, w~~We are able to find some benefits of our regional model over ~~CT2016~~our global model when looking at the diurnal cycle of CO_2 concentrations at particular sites in which large topography mismatches exist, e.g., WGC. ~~CT2016 and o~~Our global models did not capture diurnal cycles well, while our regional model did. This is a promising result because it suggests that using night time data in an inversion to estimate night time fluxes (e.g. Lauvaux et al., 2008) may be beneficial if a high resolution model is used. Currently, a GHG state estimation system using GEM-MACH-GHG and ECCC's operational Ensemble Kalman filter data assimilation system (Houtekamer et al., 2014) is under development.

4. P18/L33, add Feng et al, 2016, another study demonstrates that both of the high resolution transport and fluxes are demanding for accurate CO2 simulations at the urban scale. It also links to the last paragraph of the conclusion.

Response: Following the reviewer's comment, we have added Feng et al. (2016) in the suggested places.

Feng, S., Lauvaux, T., Newman, S., Rao, P., Ahmadov, R., Deng, A., Díaz-Isaac, L. I., Duren, R. M., Fischer, M. L., Gerbig, C., Gurney, K. R., Huang, J., Jeong, S., Li, Z., Miller, C. E., O'Keeffe, D., Patarasuk, R., Sander, S. P., Song, Y., Wong, K. W., and Yung, Y. L.: Los Angeles megacity: a high-resolution land–atmosphere modelling system for urban CO_2 emissions, Atmos. Chem. Phys., 16, 9019–9045, https://doi.org/10.5194/acp-16-9019-2016, 2016.

5. P19/L7, add Díaz-Isaac et al., 2018, which demonstrate that the meteorological IC/LBC is one of the big players in regional CO2 simulations.

Response: Following the reviewer's comment, we have added Diaz-Isaac et al. (2018) in the suggested place.

Page 19, lines 28-29: For example, the meteorological IC and LBC contribute to the variability of daytime CO_2 in the PBL (Díaz-Isaac et al., 2018).

Díaz-Isaac, L. I., Lauvaux, T., and Davis, K. J.: Impact of physical parameterizations and initial conditions on simulated atmospheric transport and CO_2 mole fractions in the US Midwest, Atmos. Chem. Phys., 18, 14813–14835, https://doi.org/10.5194/acp-18-14813-2018, 2018.

6. Figure 1, add the codes of the observation sites.

Response: Following the reviewer's suggestion, we have added codes in the figure.

7. Figure 7, it's a very dense figure. The panels should be enlarged. To make the results stand out, the authors should use lines instead of markers for presentation.

Response:
As the reviewer requested, we have made Figure 7 clearer by enlarging panels and thickening lines of error bar. As a result, new Figure 7 is now visibly clearer. But, we found that lines look sometimes too messy so we did not replace markers by lines.

8. Figure 8, remove "Should replace STD by STDE" in the caption.

Response: We have removed unnecessary words in the caption.

9. Figure 12, replace the greens with another color. The green numbers are not distinguishable in most of the cases. The figures should be enlarged.

Response: The green numbers were replaced by darker colours to enhance the visibility those numbers. The size of figure was also enlarged.

10. Figure 14, can the author extend the time scale from 92 D to at least half year? It should be very interesting results from hourly to a half year time scale.

Response: Due to the nature of the method we used for extracting the amplitude of signals from hourly to a longer time scales, any gaps in data is not allowed. We checked the availability of observation data again. LEF has 10-month period hourly data from March to December 2015, and WGC has 4-month period hourly data from May to August 2015, both at their highest level of intake height. We applied the same method to data for a longer period, separately (Figure R4). Some differences between the original (Fig. 14) and revised figure (Fig. R4) are apparent particularly on the longer timescales. This makes sense because such timescales would need much longer time series for robust results. However, our main conclusions still hold. The LAM still captures better synoptic and diurnal amplitudes than the GLB90 and GLB45 on both the timescales shorter and longer than 92 days. Therefore, we decided the keep the figure 14 in order to keep the structure of analysis throughout figures. Instead, we added Figure R4 in the supplementary as Figure S2 for those who are interested in looking at the result for a longer period.

Page 16, lines 8-10: Hence, the larger mismatch of topography results not only in inaccurate daily time scales but also other scales such as synoptic scales longer than 4-days. A similar result was also found for time periods from 92 to 300 days (Fig. S2).

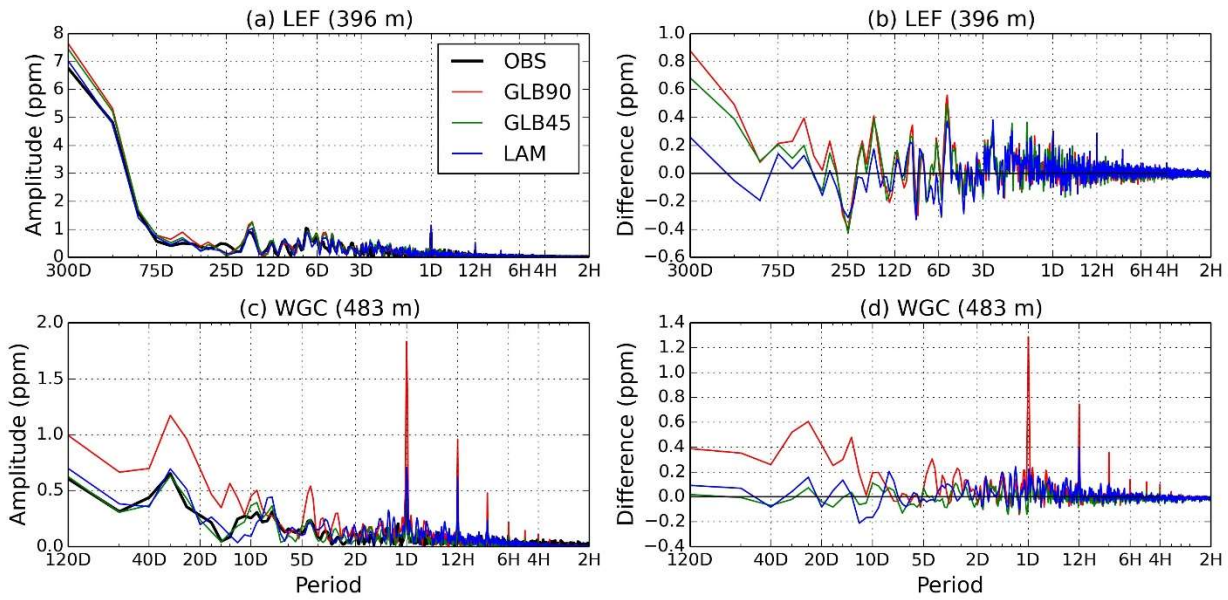

Figure R4: The amplitude of hourly time series of observed CO_2 (black) and modelled CO_2 concentrations from GLB90 (red), GLB45 (green) and LAM (blue) experiments across temporal scales from 2 h to 300 days from March to December 2015 at (a) LEF (the intake height at 396 m) and (b) their differences, from 2h to 120 days from May to August 2015 at (c) WGC (intake height at 483 m) and (d) their difference.

11. Figure 15, add "in" following "zoomed" in the caption.

Response: Following the reviewer's comment. We have corrected the caption. Note that Figure 15 is now Figure S3 in the supplementary as explained in the specific comment #2 above.