We would like to thank the reviewer for the valuable comments and suggestions for improving our manuscript. Please find below a point-by-point response (in blue) to each of the comments raised by the reviewer (in black). All the figure numbers correspond to the revised manuscript.

Anonymous Referee #3

Received and published: 7 August 2020

General Comments:

This study by Lian et al., 2020 attempts to identify and quantify significant sources of errors that can hinder the accurate estimation of urban-scale emissions. The objective of this study, as claimed by the authors, also includes demonstrating how these diagnostics can be used for inverse modelling studies. An ensemble of WRF-Chem simulations are performed, varying emission inventories (one month of simulations), PBL schemes and urban canopy schemes (one year of simulations), and boundary conditions (one year of simulations). The topic is fascinating and is essential to investigate the how sensitive is the emission estimate to the different components of the transport mechanisms (simulated by the model), flux variations, and assumptions/methods employed. This is a well-written manuscript with clearly described methods and results arranged in a logical order, which made the manuscript easy to follow. The conclusions drawn, based on their analyses, are reasonable and are applicable to those working on city-scale and mesoscale inversions of CO_2 using WRF-Chem.

We thank the reviewer for his work and this positive general assessment of the manuscript.

The study in this present form, however, fails to justify the title. Though the study considered an ensemble of simulations and subsequent analyses, it is still insufficient to make a quantitative estimation/evaluation of sources of errors in CO_2 simulations which is adequate to the broad spectrum of inverse modelling studies. I'd instead consider it as a study on diagnosing the effect of vertical mixing (PBL schemes), specific modelling criteria (urban schemes) as well as boundary conditions in city-scale modelling, in addition to assessing the sensitivity of simulations to the emission patterns. I believe that the title can be reworked accordingly to present the study appropriately. Additionally, some other sections require more clarification, modification, and further analysis, as mentioned below. Thus, I would recommend a major revision before considering for publication in ACP.

We greatly appreciate the reviewer's suggestion and fully agree that the title needs to be made more concise and to have a clearer description of the objective so that the reader can more easily understand the analysis undertaken. We thus have altered the title to better reflect our intent:

"Sensitivity to the sources of uncertainties in the modeling of atmospheric CO_2 concentration within and in the vicinity of Paris city"

Though it is mentioned as one of the objectives, I don't see that the study has addressed the question to what extent or how the model-measurement error can be reduced. This is a major concern of mine. I'd consider that the authors could devise efficient analysis strategies to address this, given the availability of measurements from 6 +2 sites and ensemble of simulations. An adequate diagnosis of model-measurement mismatches is missing, which is a weakness of this manuscript. A discussion on how the diagnostic results can be used for the betterment of inversion studies is vaguely articulated in the manuscript. All the guidelines for the data selection put forward by the study (such as discard data with high model-data misfits, use only afternoon values; test the influence of boundary conditions) is already known to the community and currently practised in inverse model calculations. I would suggest authors avoid the above statement of objective or revise thoroughly while incorporating additional analysis to explain the novelty of their findings.

We have expanded the discussion section and attempted to use it to consolidate the conclusions as far as possible. The following two paragraphs have been added to provide the perspective on reducing the model-measurement error that would be beneficial to the future atmospheric inversion studies.

"Furthermore, it remains difficult to interpret and use quantitatively in situ measurements within the city as long as there is no proper information about the turbulent airflow within and above the urban canopy. The near-surface mixing is not only controlled by the atmospheric stability conditions but also affected by the urban roughness and anthropogenic heat production. If the complex vertical mixing processes cannot be properly constrained in the transport model, it will be difficult to use the measurements acquired close to the sources in the atmospheric inversion system. Therefore, regular measurements of vertical CO_2 profiles, combined with relevant upper-air meteorological data (e.g., potential temperature and wind) and the mixing layer heights in the lower troposphere are expected to be included in the future Parisian CO_2 monitoring network. Such complementary measurements will be of great help to understand the characteristics of CO_2 vertical distribution under both stable and convective boundary-layer conditions. It can also be used to verify and validate the atmospheric transport model, and to reduce transport errors based on the data assimilation of more meteorological observations, leading to much higher accuracy in the atmospheric inversion system that aims at retrieving urban CO_2 fluxes."

"Focusing on the Paris region, two limitations of this study should be acknowledged and worth further investigating based on the high-resolution urban ecosystem modeling and monitoring so as to better quantify the impact of urban biogenic fluxes: (i) due to the coarse-resolution (1 km) SYNMAP land use data used for the VPRM model, the simulated biogenic fluxes in center Paris in this study are almost zero except for a few grid cells containing two big parks that are located in the eastern and western outskirts of the Paris city. While in reality, there are still a number of green space and pervious landscaped areas unevenly distributed in the city of Paris that need to be considered with a fine-scale (sub-kilometer) model; (ii) there is a lack of validation of the Paris-VPRM model in this study since no eddy covariance measurement is available within the Paris urban area and its surroundings. This limitation could be overcome by an expansion of the observation network with the neighborhood-scale urban eddy covariance flux measurements included."

Specific Comments:

Fig. 2. For Line style descriptions, please use another colour (e.g. black) than those used as line colours. Blue is already used for "Total". I had a hard time understanding. Also, I'd suggest you remove (c) and (d) and include AirParif (daily) and Constant in (a) and (b).

Figure 2 is changed as suggested.

Differences between CAMS and CarbonTracker at the four lateral boundaries: Why are there substantial differences between daytime and nighttime in East boundaries, sometimes even in opposite phases (Fig. 3b)? Please explain.

The CAMS and CarbonTracker could be considered as two independent global CO_2 analyses and atmospheric inversion systems. They have used a variety of different inputs (e.g. different prior estimates of CO_2 fluxes, observation datasets, etc.) and approaches or tools (e.g. different global atmospheric transport models with different meteorological forcing, different data assimilation methods with from the variational method applied to long temporal window and an ensemble Kalman filter-based method applied to much shorter analysis windows) to estimate the CO_2 surface fluxes and transport them to simulate the three-dimensional atmospheric CO_2 mole fraction. Therefore, it is hard to fully rule out the detailed causes of the substantial difference at the eastern boundary. A possible explanation is the larger differences between both the fossil fuel and biogenic CO_2 fluxes over the European continent as a result of their respective analysis (http://www.globalcarbonatlas.org/en/CO2-emissions). It may also because of the sensitivity of the modeled CO_2 concentrations to the transport fields over the mountain region in the eastern part of the outer domain.

We have added the following sentence in the manuscript to account for the reviewer's comment:

"A possible explanation could be that both fossil fuel and biogenic CO_2 fluxes and associated uncertainties are larger over the European continent than over the oceans. It may also be caused by the sensitivity of the modeled CO_2 concentrations to the transport fields over the Alps mountain region at the eastern boundary."

Evaluation with in situ observations: I am not very convinced with the usage of KNN method? How is the outlier fraction calculated? How sensitive is the filter size in another outlier fraction? What are the criteria for the choice of 0.1 in this case?

Regarding the evaluation in section 3.1, it is worth noting that we also look at statistics without removing outliers, and the text is based on the analyses both with and without the KNN method. Moreover, please see our answer to Reviewer #1 on this topic where we show indications that this method successfully removes outliers that are due to the measurement contaminations from local unresolved sources of emissions and/or the model's inability under complex meteorological conditions. Regarding the conclusion/discussion section, we recognize that it would be more appropriate to encourage the use of this KNN algorithm based on a deeper analysis of the detected outliers instead of using it as a crude data filtering. We thus have rephrased the text in the revised manuscript (also see answer to Reviewer #1 for details).

The basic idea is that errors from emissions or from patterns that can be modeled at 1 km resolution should have a frequency and timescale longer than that of misfits removed by the algorithm. The algorithm is expected to remove isolated outliers that are not representative of normal conditions. It remains an arbitrary issue in terms of the definition of an outlier and the decision of whether to remove or keep it.

In response to the reviewer's concerns, we made a test of sensitivity to the outlier fraction, with values for this fraction ranging from 0 to 0.3. Figure R3 shows the change of the RMSE and MBE as a function of the outlier fraction values at one urban station (CDS) and one suburb station (SAC). In this study, the choice of 0.1 was an arbitrary selection based on the visual judgment of Figure R3 and Figure 5 (scatter plot of the model-data comparison).

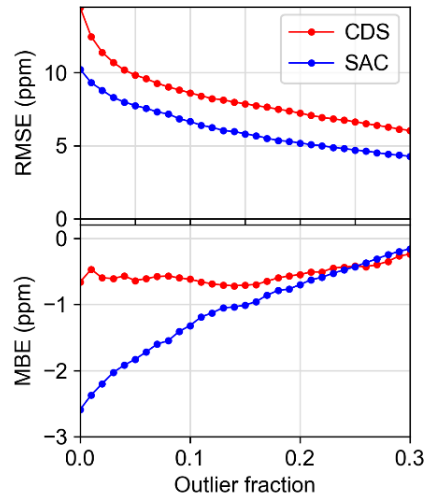

Figure R3. Change of the RMSE and MBE (statistics for all hourly model-data CO_2 comparison from December 2015 to November 2016) as a function of the outlier fraction values used in the KNN method at one urban station (CDS) and one suburb station (SAC).

Given that outliers are removed, why the model-observation mismatch is this high (Fig. 4)? How can these mismatches be reduced? How about background sites in terms of evaluation? In addition to reporting the mismatches, I'd suggest the authors explain the reasons for these large deviations from observations. This is critical as I also see unexpectedly significant model-measurement differences in diurnal averages. Have authors checked different choices of physics/dynamics schemes or other parameters available in WRF-Chem to reduce this mismatch?

Note: Figure 4 is now ranked as Figure 5 in the revised manuscript.

This large all hourly model-data mismatch is mainly because the model underestimates CO_2 with a bias ranging from 0 to 12 ppm across stations during the night until around 05 UTC (also see Figure 6). In fact, a fairly detailed explanation of the two main causes of this model-data discrepancy was already provided and justified in section 3.2 of the manuscript. It is probably due to the prescribed nighttime heating emission profile used in the AirParif anthropogenic inventory (the second paragraph in section 3.2.1) and the nighttime model transport issue (the third paragraph in section 3.2.1). More precisely, our results first indicate that the temporal profile of the heating sector used by the AirParif inventory (with a significant decrease along the night) tends to bear a large uncertainty. In the IdF region, CO_2 emissions from the heating sector are linked to the burning of gas and oil, and electricity consumption. We could expect that a more constant diurnal profile should probably be a better approximation to the truth than the current one. This uncertainty could be justified and reduced by a further analysis of these related source data. Moreover, another two paragraphs in terms of the atmospheric transport and the modeled biogenic flux have also been added in the discussion section to provide the perspective on reducing the model-measurement error (please see our answer above).

We have also checked the background site at TRN station, located 101 km away from the center of Paris (see Figure 1) with air inlets placed at 5m, 50m, 100m and 180m AGL respectively. Figure R4a shows the scatter plot of the observed and BEP_MYJ (control run) simulated all hourly CO_2 concentrations at 50m and 100m sampling heights from December 2015 to November 2016. The comparison of the average diurnal variations between observation and model is shown in Figure R4b. Results show that the model also underestimates the CO_2 concentrations at this background site during the night, which is consistent with the conclusions obtained from the stations within the IdF region. It is worth noting that both the observation and the model show an increase of CO_2 during the night at TRN, whereas the opposite was found to be true for the modeled value within IdF. This is because the fossil fuel CO_2 emissions outside the IdF region were taken from the IER inventory that has a more constant nighttime diurnal profile than the AirParif does.

We also made the model-data CO_2 comparisons of the other four sensitivity tests of physics schemes in this study (Table 1a). The results are shown to be similar to the control run (BEP_MYJ), except that the model runs with the UCM scheme tend to have larger misfits at the two urban stations (JUS and CDS). The two schemes differ in their representations of the near-surface mixing which leads to large differences in the modeled CO_2 concentrations (please see our answer to Reviewer #2 for details). On the other hand, we also evaluated WRF simulated meteorological fields against the observations. Results indicate that the model does not suffer from obviously inappropriate/wrong parameter settings that could lead to a significant transport error (also see our answer to Reviewer #2 for details).

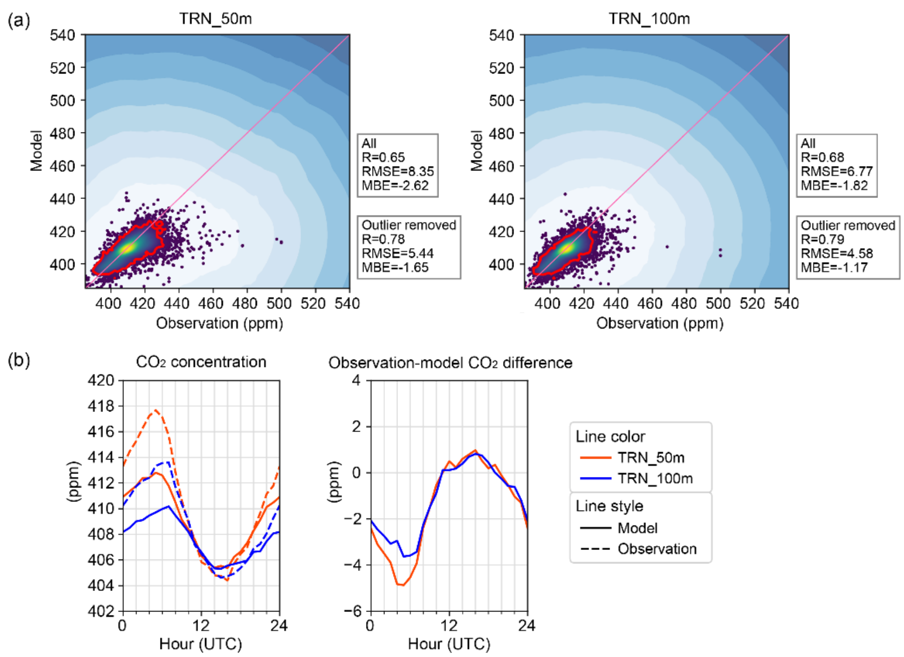

Figure R4. (a) Observed and BEP_MYJ (control run) simulated all hourly CO_2 concentrations at TRN site with sampling heights at 50m and 100m above ground level respectively from December 2015 to November 2016. (b) Comparison of the average diurnal variations between the observed (dash line) and modeled (solid line) CO_2 concentrations, and the model-data CO_2 misfits at TRN site.

Sect. 3.2.1 "modeled value (green line in Figure 6) gets somewhat closer to the observation" In Fig. 6, the blue curve represents constant emissions. Please check. The simulations (BEP_MYJ_CON) reproduce the observed patterns better than other simulations; however, not in magnitude. So please rephrase the sentence accordingly: "modeled value (green line in Figure 6) gets somewhat closer to the observation".

Note: Figure 6 is now ranked as Figure 7 in the revised manuscript.

As described in section 2.1.2, the green line (BEP_MYJ_AIP) is the AirParif inventory with a constant temporal profile (each pixel has a different emission, but constant in time based on the temporal average of the AirParif inventory). The blue line (BEP_MYJ_CON) is a constant and spatially homogeneous emission where the emissions are distributed uniformly over the IdF whole territory.

The original sentence in section 3.2.1 is "Indeed, when assuming an emission constant in time, the decreasing trend is reduced and the modeled value (green line in Figure 7) gets somewhat closer to the observation."

For better clarity, we have changed the "an emission constant in time" to "the AirParif inventory with a constant temporal profile". The modified text is as follows:

"Instead, the decrease of anthropogenic emissions during the night (Figure 2) explains part of the decrease in modelled concentrations. Assuming the AirParif inventory with a constant temporal profile, the decreasing trend at night is reduced and the modeled value (green line in Figure 7) is closer to the observation than the control run (BEP_MYJ)."

Also please indicate the season (or January) in the Figure 6 caption.

Note: Figure 7 is now ranked as Figure 8 in the revised manuscript.

Done

Also, I am happy to note that the authors demonstrate the effect of emission trend and atmospheric vertical mixing here. Please comment on the effect of boundary conditions (though I expect it to be minimal by looking at the patterns in BEP_MYJ_CON).

We have updated Figure 7 to include the average diurnal cycle for the biogenic and background CO_2 along with the one of the total and anthropogenic CO_2 concentrations. It is worth noting that only the control run (BEP_MYJ) was plotted. This is because there were 6 distinct CO_2 tracers within this one-month simulation. Four of them corresponded to each of the four sensitivity tests of anthropogenic emissions to record their distinct atmospheric CO_2 signatures. The other two tracers were for the biogenic and background CO_2 respectively. Therefore, there will be no difference in the simulated biogenic and background CO_2 values among the four sensitivity tests. Results in Figure 7c and 7d show that the impacts of biogenic flux and background condition on this simulated decrease are relatively small, which is on the order of a fraction of a ppm.

The revised manuscript has included the following sentence as suggested by the reviewer:

"Results in Figure 7c and 7d show that the impacts of biogenic flux and background condition on this simulated decrease are relatively small as they are on the order of a fraction of a ppm."

PBLH and vertical distribution of the modelled CO_2 (BEP_MYJ): It is not clear to me why authors have mentioned Nielsen-Gammon et al., 2008 for PBLH estimation. What extent the MYJ scheme and Nielsen-Gammon et al., 2008 differ in deriving the PBLH? If different, a comparison plot will be helpful here.

Numerous thermodynamic parameters, including temperature, humidity and their derivatives (e.g., potential/virtual potential temperature) have been widely used to define the PBL height. The 1.5-theta-increase method defines boundary layer top based on the potential temperature and it has widely proven to be fairly representative and practical for PBL height detection. The diagnosed PBL heights obtained directly from different PBL schemes in WRF are calculated based on their own specific formulations. The MYJ scheme determines the PBL height using the turbulent kinetic energy (TKE) profile. The top of the PBL is defined to the height where the TKE profile decreases to a prescribed low threshold value (Janjic 2001). For the purpose of a fair comparison among the 3 PBL schemes (YSU, MYJ, BouLac) tested in this study, the 1.5-theta-increase method was used as a criterion to diagnose PBL heights. However, these comparison results are not shown in Figure 8 as they might be fruitless for the objective here.

In response to the reviewer's concerns, we performed a comparison of PBL heights estimated by the model against measurements at SIRTA station located about 20 km southwest of Paris center during January 2016 (Figure R5). Generally, the temporal variation of the diagnosed PBL heights is shown to be similar for the MYJ method and the 1.5-theta-increase (NG) method. The correlations with hourly observations are similar for MYJ (0.57) and NG (0.56), but the associated RMSE and MBE are better for NG. Both methods indicate that the model control run (BEP-MYJ) overestimates the PBL heights with a MBE of 136 m for MYJ and 54 m for NG.

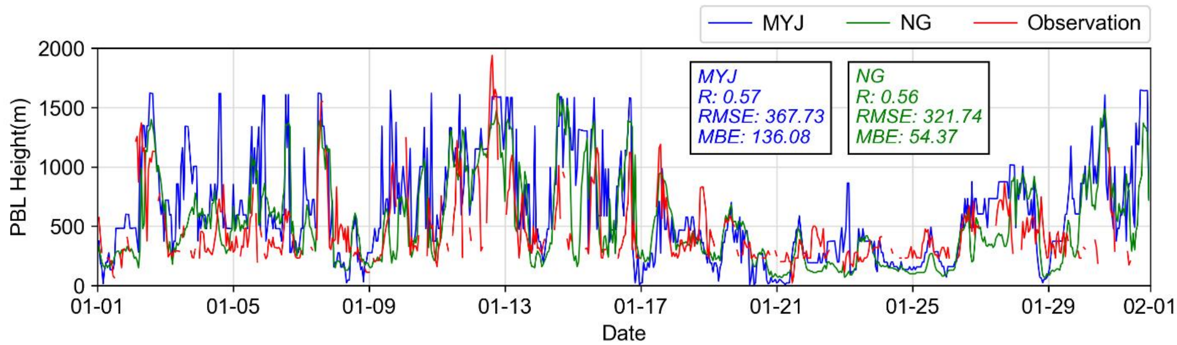

Figure R5. Time series of the BEP-MYJ (control run) simulated PBL heights diagnosed by the MYJ scheme (in blue) and the 1.5-theta-increase method (NG, in green) with respect to the observations at SIRTA station (in red) for January 2016.

I am a bit surprised with the high PBL values in winter (initial half-month) over Paris. Do authors look at the PBL measurements (e.g. lidar measurements)? Fig. 7 is confusing as the 34-m AGL curves have nothing to do with the left Y-axis values. I would suggest authors separate these two curves (magenta and pink) from this and make an independent plot along with PBLH.

Note: Figure 7 is now ranked as Figure 8 in the revised manuscript.

Figure 8 has been modified as suggested. The time series of the wind arrow and the model-data CO_2 concentration has been moved to a new panel (Figure 8b). Apart from changing the line color, we also added the PBL height measurements in Figure 8a. The PBL heights data were obtained at the SIRTA station located about 20 km southwest of Paris center (Kotthaus et al., 2020. https://sirta.ipsl.fr/) as shown in Figure 1 in the revised manuscript. The comparison of all hourly measured and modeled PBL heights show that the model could well reproduce the temporal variation of the PBL heights in January and there is not an obvious long-term bias.

In addition, we also carried out a one-year validation of the PBL heights. Results in Figure 4d and Figure S2d further confirm that the model is capable of reproducing the PBL heights for other months in the year.

Sect. 4: Please see my comments above (w.r.t title) and refine this section thoroughly.

Please see the answer above.

Minor comments:

Page 7, "from 18 pm to 22 pm)" Please change to 18:00 UTC to 22:00 UTC.

Text is changed as suggested.

All references mentioned in this response are already included in the manuscript.