Review of Lian et al., "Quantitative evaluation of the uncertainty sources for the modeling of atmospheric CO2 concentration within and in the vicinity of Paris city"

General comments: The authors attempted to determine the uncertainty sources for the modeled CO2 concentrations over Paris, France, using a set of WRF-Chem simulations varying with physics-based transport, fossil fuel emissions, and CO2 boundary conditions, for 2016. They mainly focused on the impact of PBL schemes and with the combination of the urban canopy models, two fossil fuel emission inventories with /without hourly variability, and two global models as boundary conditions on the modeled CO2 in comparison with the ground-based in-situ CO2 measurements. Their results show that model-data mismatch maximizes in the nighttime so they recommended the readers to discard the model-data misfits and use afternoon measurements for inversion. This is not new, and I believe that is what we do in atmospheric inversion. They also found the boundary condition could cause large differences at the synoptic scale and suggest exploring more about the influence of the boundary condition on the inversed results and suggest using additional observation to constrain boundary conditions. This is also not new. The authors are aware of these points because they cited those papers. So, I failed to locate the novelty of the work that brought into the community. The authors, in my opinion, have repeated some of the previous studies without extending the science further.

Besides, the authors cited a few pilot CO2 urban studies, such as INFLUX and LA megacity. Both Lauvaux (2016) and Feng (2016) pointed out the significant improvement of using high-resolution fossil fuel inventories in simulating CO2 at the urban environment. Although the authors included two different fossil fuel inventories in the simulation with the variation of the temporal components, I have a hard time following the goal of the experimental design. I thought they would explore the sensitivity of modeled CO2 to the temporal resolutions when I was reading the methods and section 3.2.1, but the related findings were not emphasized in the conclusion. Why?

One of the major concerns in the urban CO2 studies falls in the impact of the biosphere around or within the cities. The results of this study also showed that the impact of the biogenic fluxes is significant and not negligible, meaning that the biosphere is another uncertainty source over Paris. The author can refer to Feng (2019a; 2019b) to construct a set of biospheric fluxes and investigate the uncertainty of the biospheric in the modeled CO2 as well. Additionally, the authors relied on the VPRM module in WRF-Chem to provide the biogenic fluxes. It's not clear to me that if VPRM has been tuned with flux towers or not. The VPRM parameters in WRF-Chem are fixed and needed to be tuned with the flux towers to have a relatively accurate biospheric flux estimation (Hilton et al., 2013: Hilton et al., 2014). If the authors used the default values for the VPRM parameters in this study, the authors will have to consider the errors caused by the biosphere during the interpretation of the results which is almost impossible to isolate from transport, emission, or boundary condition. However, because of the simplicity of VPRM, the authors can build a set of parameter-based perturbations of the biospheric fluxes via VPRM to address my first concern.

The authors chose five combinations of the PBL schemes and urban canopy models to study the impact of the transport and concluded that it's difficult to have "good" transport. First of all, what are the rationales the authors believe these tow schemes are the key players of the CO2

urban modeling? Díaz-Isaac (2018) using an ensemble approach pointed out PBL indeed is a major player but ranked No. 2. The most dominant parameterization is the land surface model used in the study. I am aware that the response of the modeled CO2 to the model physics may vary when the location changes. Have the authors explored the impact of other model parameterizations on the simulated CO2? This may be also why the model results have such a large bias in this study. Secondly, the model-data mismatches are extremely large and out of my expiation. For example, the whole year averaged diurnal mismatch can be as large as -10 to 5 ppm at the two urban sites in Figure 5. I found a similar figure in Figure 9 of Feng (2016), even though it's a month averaged value, in which the diurnal cycle from the high-resolution simulation looks almost identical to the observation. What causes the large bias in this work? I would check if any errors caused by other model physics. Thirdly, the authors concluded that the transport issue is difficult to identify. I disagree. The model transport can be evaluated with meteorological observations. Apparently, there are meteorological observations at the monitoring sites. Additionally, there are quite a few WMO stations in the domain. Comparing with meteorological observations in the model domain will allow the authors to have a better sense of the model transport.

As the authors mentioned that boundary conditions can lead to large bias in the inversed results, the results showed that 5-20 ppm day-to-day difference between the two global models along the edges of the model domain. In the CO2 regional (inverse) modeling, one of the major concerns about the boundary condition is the conservation of mass (Butler et al., 2020). How did the authors handle mass conservation when incorporating global modeled CO2 into the regional model domain? Another issue is that the number of boundary conditions used is too small to quantify the uncertainty. Strictly speaking, to be able to claim quantification of the uncertainty sources, a large number of the ensemble and a set of calibration procedures are required, such as rank histogram, reliability diagram, brier scores, etc. (Garaud and Mallet, 2011). Although it may be difficult to meet two criteria with the CO2 modeling, the authors will at least need three of them to study the sensitivity.

In summary, this work claims that it has a quantitative evaluation of uncertainty sources in the CO2 modeling, but the experimental design is far from achieving the goal. It eventually is merely a sensitive study of modeled CO2 to the selected fossil fuel emissions, the combination of PBL and urban canopy models, and boundary condition. The size of the ensemble they built does not allow them to do a solid quantification study. As I mentioned, this study appears repeating some of the previous studies without advancing the understanding the community already holds currently, neither in science nor in techniques.

There are no clear rationales why they made such selection as I pointed out with the transport "ensemble". The authors did not address the major issues in urban modeling, i.e., the impact of biosphere, and regional modeling, i.e., the conservation of mass when applying boundary conditions. They also failed to have a clear conclusion about the findings associated with fossil fuel emissions. In my opinion, this work is incomplete and must be extended to consider publication; these concerns I brought up can be addressed, which, however, will require a new design of the method. In addition to the specific comments I listed below, I would not recommend this MS to be a published in ACP.

Specific comments:

Section 2: There are important details missing in the description of the model setup.

1) Did the model use simulation cycles? If yes, how often is it? If yes, how was the CO2 field addressed, initializing every time or being carried over simulation cycles?
2) ERA-Intrim and the outermost domain of WRF-Chem have quite different resolutions. What are the rationales that the authors used grid nudging over spectral nudging?
3) As I mentioned in the general comments, has the VPRM parameters constrained with the flux tower measurements?
4) When the authors were incorporating CO2 IC/BC to WRF-Chem, how did the author address the conservation of CO2 mass?
5) When using global modeled CO2 as IC, the discontinuity of the global and regional model dynamic can cause discrepancy of the CO2 as well. How much the difference caused by the discontinuity would be?

P 6, L 25-30: the author interoperated that the reason of the higher CO2 concentrations in Fall than in winter was due to the anticyclone keeping the high CO2 in the domain for quite a while. I disagree. If it's due to meteorology, the impact on the fossil fuel CO2 and biospheric CO2 concentration should be the same. We should see lower CO2 in the suburban sites, but we don't.

P7, L1-5: as I said earlier, the authors should be able to identify at least to some degree if the issues are in transport or boundary conditions by comparing with the meteo data.

P10, L10-15: what causes the different bias between the BEP and UCM schemes? I would like to see a deeper explanation of that instead of simply saying lower or higher.

P11, L19-21: I agree that based on the current setup, there is little hope to improve the model performance. However, the authors can follow my suggestion listed in the general comments. For example, checking the land surface model used, comparing with meteo data, etc., to identify if the problem is caused by transport is the first step. Then the authors can look into the emission, boundary conditions, etc.

Figure 5: please use local time in the x-axis instead of UTC. The much bigger issue is the large bias in the biases.

Reference:

Butler, Martha P., Thomas Lauvaux, Sha Feng, Junjie Liu, Kevin W. Bowman, and Kenneth J. Davis. "Atmospheric Simulations of Total Column CO2 Mole Fractions from Global to Mesoscale within the Carbon Monitoring System Flux Inversion Framework." Atmosphere 11, no. 8 (August 2020): 787. https://doi.org/10.3390/atmos11080787.

Díaz-Isaac, Liza I., Thomas Lauvaux, and Kenneth J. Davis. “Impact of Physical Parameterizations and Initial Conditions on Simulated Atmospheric Transport and CO2 Mole Fractions in the US Midwest.” Atmospheric Chemistry and Physics 18, no. 20 (October 16, 2018): 14813–35. https://doi.org/10.5194/acp-18-14813-2018.

Feng, S., Lauvaux, T., Newman, S., Rao, P., Ahmadov, R., Deng, A., et al. (2016). Los Angeles megacity: a high-resolution land–atmosphere modelling system for urban CO2 emissions. Atmospheric Chemistry and Physics, 16(14), 9019–9045. https://doi.org/10.5194/acp-16-9019-2016

Feng, Sha, Thomas Lauvaux, Kenneth J. Davis, Klaus Keller, Yu Zhou, Christopher Williams, Andrew E. Schuh, Junjie Liu, and Ian Baker. “Seasonal Characteristics of Model Uncertainties From Biogenic Fluxes, Transport, and Large-Scale Boundary Inflow in Atmospheric CO2 Simulations Over North America.” Journal of Geophysical Research: Atmospheres 124, no. 24 (2019): 14325–46. https://doi.org/10.1029/2019JD031165.

Feng, Sha, Thomas Lauvaux, Klaus Keller, Kenneth J. Davis, Peter Rayner, Tomohiro Oda, and Kevin R. Gurney. “A Road Map for Improving the Treatment of Uncertainties in High-Resolution Regional Carbon Flux Inverse Estimates.” Geophysical Research Letters 46, no. 22 (2019): 13461–69. https://doi.org/10.1029/2019GL082987.

Garaud, D., and V. Mallet. “Automatic Calibration of an Ensemble for Uncertainty Estimation and Probabilistic Forecast: Application to Air Quality.” Journal of Geophysical Research: Atmospheres 116, no. D19 (October 16, 2011). https://doi.org/10.1029/2011JD015780.

Hilton, T.W., K. J. Davis, and K. Keller. 2014. Evaluating terrestrial CO2 flux diagnoses and uncertainties from a simple land surface model and its residuals, Biogeosciences, 11, 217-235, doi:10.5194/bg-11-217-2014.

Hilton, T.W., K. J. Davis, K. Keller, and N.M. Urban. 2013. Improving North American terrestrial CO2 flux diagnosis using spatial structure in land surface model residuals, Biogeosciences, 10,4607–4625, doi:10.5194/bg-10-4607-2013.