We would like to thank Referee #1 for his/her thoughtful comments and detailed suggestions to our manuscript. In the following, we answer to the reviewer's comments and indicate the changes in the manuscript that were implemented according to the recommendations. The comments are in black. Our answers are in blue. All the figure numbers correspond to the revised manuscript.

Anonymous Referee #1

Received and published: 31 July 2020

General Comments

The study investigates potential sources of error for estimating urban CO_2 emissions using atmospheric observations. Understanding and mitigating these errors is necessary to produce more accurate emission estimates, and the authors suggest several criteria to select data to avoid the impact of these error sources. While the study focuses on Paris, the methods and results presented are more widely applicable and of interest to other urban emission estimation schemes using atmospheric data and transport models.

The methods of the paper focus on examining a set of clearly described WRF-Chem forward model runs, with the CO_2 emissions, boundary conditions and physics schemes varied. These comprise a logical set of factors to explore, with the authors acknowledging this is a subset of all possible error sources but is still shown to be important. The detailed analysis of these results links well with the corresponding conclusions drawn (suggestions ii and iii of the abstract). I believe this paper is a useful contribution to the field, providing quantifications of significant sources of uncertainty in current models and providing a framework for further urban systems to examine their own uncertainties. However, I do have a few concerns with other aspects of the paper, as detailed below. Therefore, I recommend this paper for publication in ACP once the issues outlined below have been addressed.

We thank the reviewer for these very supportive comments.

Specific Comments

Introduction - The study could be seen as an extension to previous works in looking at sources of uncertainty (Martin et al 2019 https://doi.org/10.1016/j.atmosenv.2018.11.013) and are complementary to other recent studies on uncertainties in estimating urban emissions (such as Balashov et al 2020 https://doi.org/10.5194/acp-20-4545-2020). The context set out in the paper could be improved by including comparisons to such other studies.

Please see our answer to Referee #2. Table R1 shows a comparison between this study and few pilot CO_2 urban studies with the objective to investigate in detail the sources of uncertainty/error in the atmospheric CO_2 modeling for cities, such as Los Angeles (Feng et al., 2016) and Washington DC/Baltimore (Martin et al. 2019). Given that the sources and characteristics of urban fluxes of CH_4 is different from those of CO_2, we have not included a comparison with the Balashov et al. 2020 here.

Page 1 line 28 – Value quoted is for scope 2 emissions, but inversions only estimate scope 1 emissions. Either this should be made clear or the authors should use scope 1 emissions value.

As suggested, we have used the value of direct emissions (scope 1). The modified text is as follows:

"cities directly release about 44 % of the global energy-related CO_2 emissions"

Page 6 line 6 – The use of the KNN outlier removal needs greater justification and is my greatest concern with this paper. The authors claim that this algorithm removes observations of sources too local to be resolved or meteorological conditions that model is less skilled with (which are valid reasons for removing data points) but provides no evidence that this is the case. As it is, the algorithm may just be arbitrarily throwing away data that highlights systematic over or underestimates in the emissions field that is needed for an inversion. Either it needs to be demonstrated that the algorithm only removes points that are linked to these conditions, or a different method, preferably based on physical reasoning, should be used.

In response to the reviewer's concerns about the KNN outlier removal, we did perform some further analyses and validations of this method to support the approach and the related statements in the manuscript. These analyses definitely show that the outliers generally correspond to either:

1) the model's inability under specific meteorological conditions.

After analyzing the dates of the identified outliers, we found clusters of outliers that occur as the result of weather episodes with a duration of one-to-few days. Several cases were identified and described in Table S1. One sample case, presented here, shows unfavorable meteorological conditions from Jan 18th to 21st 2016. During this 4-day period, with a return of the winter anticyclonic conditions over the entire region, dense fog and weak winds were observed. Stubborn low clouds kept temperatures chilly with little snow. Figure S3a shows the time series of the observed and modeled (using MYJ_BEP) hourly CO_2 concentration at SAC station. The grey shaded areas indicate the ranges of model results with five physical parameterization schemes used in this study (Table 1a in the manuscript). The yellow vertical lines indicate the large model-observation misfits (outliers) detected by the KNN algorithm. It can be seen that for the certain hours that were tagged as outliers, the differences between observed and modeled CO_2 concentrations can be as large as 70 ppm. Meanwhile, the spread of the simulations of CO_2 is much larger than during the days before and after this period, leading to a higher mean bias error and root-mean square error of the ensemble mean. Figure S3b shows the distribution of the hourly CO_2 concentrations as a function of the wind speed for the year 2016. It clearly illustrates that the detected outliers occurred more often in weak-wind conditions (< 2.5m/s) which are difficult to reproduce by the model. From this example, we can say that KNN can detect outliers corresponding to conditions when the model physics encounters limitations.

2) the specific measurement contaminations from local unresolved sources of CO_2 emissions.

On the other hand, this KNN method was inspected for its ability to remove some CO_2 spikes due to very local influences or sampling contaminations, mainly under low wind speed conditions. We illustrate this phenomenon with the example of the measurements of hourly CO and CO_2 concentrations (CO being used to confirm the anthropogenic origin of the spikes in the atmospheric concentration) at the OVS station in 2016. The CO and CO_2 hourly mole fractions, as well as their ratios, are plotted as a function of observed wind speed and direction (Figure S4a). The location of the CRDS CO & CO_2 sampling inlet is on a building roof, where there is a building ventilation exhaust shown in Figure S4b. Figure S4a shows that the CO signal tends to be larger relative to that of CO_2 with low winds (< 4m/s) blowing from the east. This corresponds to the position of the building exhaust air system relative to that of the sampling inlet, and this is at odd with the North East position of the Paris urban area or of the main neighbor and large area sources relative to the OVS site. Further investigation shows that these CO spikes at OVS are mostly measured at night in winter, leading to a nighttime mean concentration even much larger than those two urban stations (JUS and CDS). We thus highly suspect that the measurements of CO and CO_2 are contaminated by the exhaust air of the building under specific conditions (winter nighttime with light winds). Most of the dates corresponding to these CO and CO_2 spikes exactly coincide with the outliers at OVS that have been detected by the KNN algorithm shown in Figure 5 in the manuscript. From this example, we can say that KNN can detect outliers (in the data) corresponding to real physical local contaminations.

Therefore, the KNN method, as shown above, can detect misfits between the observations and the models that would be misleading for the city scale inversions. But we also acknowledge the fact that removing data points simply based on statistical analysis without identifying the outliers on a case-by-case basis may lead to a loss of data that are suitable for the city scale inversion. In practice, manual inspection is preferable for the identification of the cause of the error. However, this is not practical given a large amount of data at six in situ stations collected over one year as those analyzed in this study. It is also difficult to find a general outlier detection method fitting to any site, model and atmospheric transport conditions.

The above analyses, table and figures will be put into the supplement. We have added the following text in section 3.1.2:

"We further analyzed the filtered hourly concentrations (detailed in supplement material Figure S3 and S4) and confirmed the contamination at one of our sites (OVS) and the relationship between meteorological conditions and excluded modeled concentrations."

Regarding the conclusion and discussion section, we agree with the reviewer that it would be more appropriate to encourage the use of the KNN algorithm based on a deeper analysis of the detected outliers instead of just saying one should use it as crude data filtering. We have rephrased the text as follows:

"We should also note that removing outliers based on statistical analysis without attributing them to a real data contamination or model limitation has potential for data loss, which could 'over-filter' the solution of an inversion for emissions. Manual inspection combined with KNN statistical filtering was shown on two examples to be a promising way to confirm that outliers have a physically justified reason to be filtered for an atmospheric inversion that aims at quantifying the city emissions. However, the amount of data removed by this filtering approach is rather low and, therefore, the information from these data should not be statistically significant for the city scale inversions. We note however that it can be critical to discard them since the least square formulation of the optimization underlying these inversions could provide much weight to these data with large discrepancies to the model."

Table S1. Meteorological conditions for several situations when large model-data misfits have been detected by the KNN algorithm.

Date	**Bulletin Climatique Météo-France***
January 19-21 2016	With the anticyclonic conditions, frosty fogs and stubborn low clouds were observed. Temperatures dropped below normal with local snow. The wind was weak to moderate.
April 12-13 2016	Disturbances crossed the region on 9th, followed by a rain-unstable rise from 10th to 13th.
August 27 2016	The weather was under some unstable intermissions, e.g., stormy on 27th and 28th, then a few showers remained on 29th.
October 25 2016	With the gradual increase of pressure until 1035 hPa on 28th, low clouds and fogs were tenacious.
…	…
* Accessible at: https://donneespubliques.meteofrance.fr/?fond=produit&id_produit=129&id_rubrique=29	

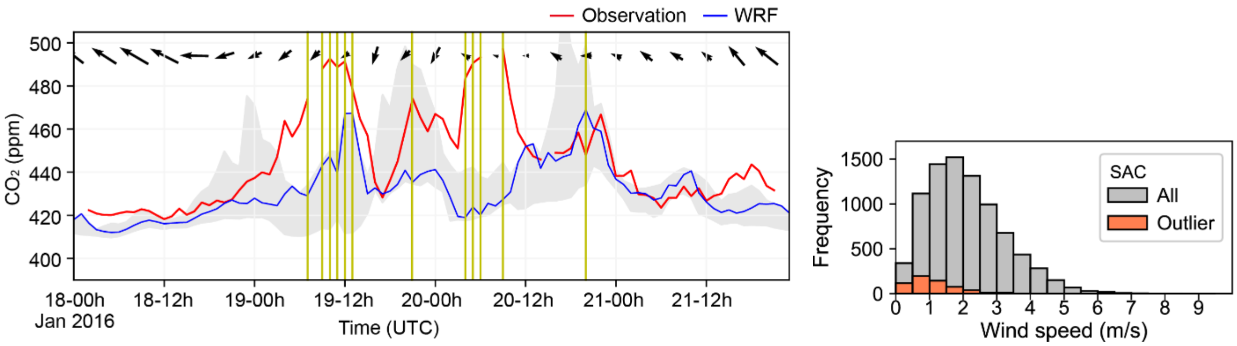

Figure S3. (left) Time series of the observed and MYJ_BEP modeled hourly CO_2 concentration at SAC station from Jan 18th to 21st 2016. The grey shaded areas indicate the ranges of simulation results with five physical schemes used in this study (Table 1a in the manuscript). The yellow vertical lines indicate the large model-observation misfits (outliers) detected by the K-nearest neighbors (KNN) algorithm. (right) Distribution of the hourly CO_2 concentrations as a function of the wind speed for the year 2016.

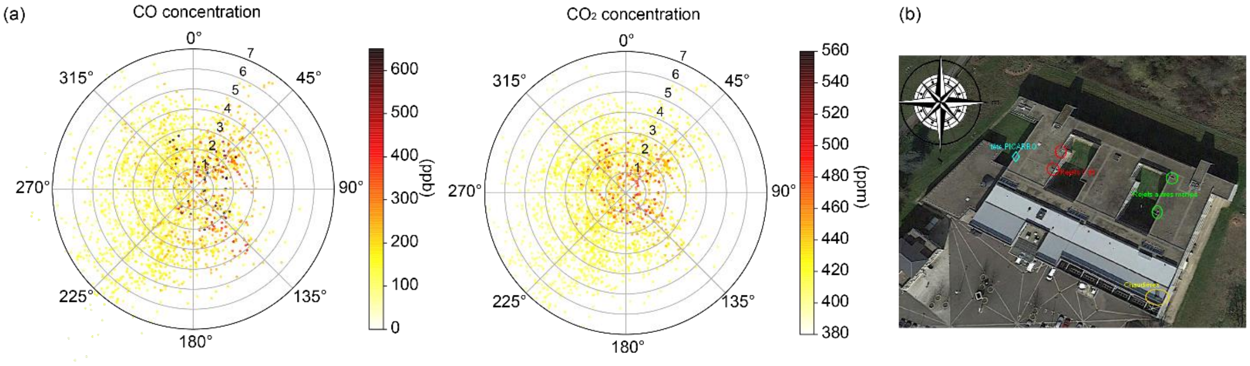

Figure S4. (a) Hourly CO and CO_2 concentration measurements as a function of wind speed and direction at OVS station for the year 2016. (b) Image of the rooftop at OVS station with the CRDS CO & CO_2 sampling inlet in cyan and the building exhaust air system in red and green.

Page 7 line 2 – The authors say that individual measurement errors are negligible compared to model-data differences. However, model-data comparisons are made on hourly time scales and there will be (potentially large) variation within the hour. How is this sub-hour variation accounted for in model-data comparison?

In response to the reviewer's question about the sub-hour CO_2 variation, we also show the distribution of the standard deviation of all minute-scale CO_2 samples per hour at CDS station for the year 2016 (Figure R1). The median value of standard deviation is around 1.3 ppm and the upper quartile (75th percentile) is less than 2.5 ppm. The magnitude of this sub-hour variability is much smaller than the RMSE value (14.5 ppm, shown in Figure 4) of model-data comparison. We thus consider that the CO_2 measurements provided by the CRDS instrument have a high precision and the individual measurement errors are negligible compared to the model-data differences. Furthermore, the model was set to output concentration at hourly scale so that we do not have the sub-hourly simulated instant concentrations to make the high-frequency model-data comparison.

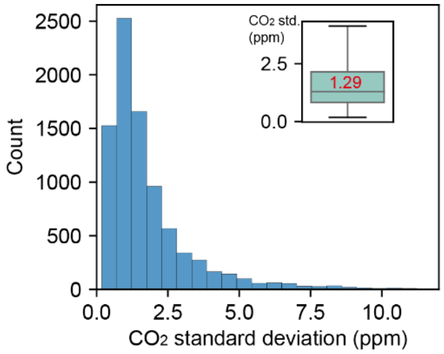

Figure R1. Distribution of the standard deviation of all minute-scale CO_2 samples per hour at CDS station for the year 2016. The midpoint, the box and the whiskers in the boxplot represent the 0.5 quantile, 0.25/0.75 quantiles, and 0.1/0.9 quantiles respectively. The median is shown as a horizontal line going through the box with the value in red.

Page 8 line 17 – The authors suggest upwind-downwind gradient can be used even in growing seasons as natural biogenic fluxes do not completely offset anthropogenic fluxes – but this still requires good knowledge of biogenic fluxes as they will make a major contribution to the observed mole fraction difference. The authors note several lines later that the biogenic fluxes show systematic errors, does this contradict the first statement?

The first statement intended to explain that biogenic fluxes in growing seasons do not offset the signal of anthropogenic emissions, but that uncertainty in biogenic fluxes makes it more difficult to quantify anthropogenic emissions in an inversion during the growing season. This is linked to the impact of biogenic fluxes on the knowledge of the anthropogenic contribution to the CO_2 gradients. We rewrote the two sentences in section 3.2.2 as follows:

The first statement as "During the afternoon, the CO_2 differences are mostly positive and result primarily from the larger contribution of the anthropogenic emissions at CDS, both during the growing and non-growing season. This result indicates that the magnitude of daytime net carbon uptake plants between the stations does not fully offset that of the anthropogenic emissions, and thus the CO_2 concentration gradients between the upwind and downwind stations that are used in previous inversion studies. Nevertheless, the biogenic contribution to the gradient is not negligible with a potential impact on the estimate of the anthropogenic emissions from the measured gradient."

The second statement as "The respiration emission at night seems to be underestimated by the VPRM model. If this nighttime respiration bias would be correlated with the daytime respiration bias (Reichstein et al., 2005), it would imply that modeled positive gradients of CO_2 between urban and rural stations could be overestimated during the growing season. We thus recommend for an inversion to control separately (with a priori) anthropogenic emissions and net ecosystem exchange, or even photosynthesis and respiration if additional data confirm a bias of respiration in VPRM."

Page 10 lines 1-21 (and Figure 11) – This section should be reworked for clarity. My understanding is that the authors have averaged the difference of tracer concentrations between runs across time (by month) and horizontal space (by land type) to calculate the individual values shows in figure 11 – but this should be made clearer in the writing. For many of the values, the standard deviation is large w.r.t. the mean difference. A different type of plot that shows the distribution, such as a violin plot, may be more appropriate.

Note: Figure 11 is now ranked as Figure 12 in the revised manuscript.

We have changed the legend of Figure 12, as well as the description at the beginning of the paragraph where Figure 12 is mentioned.

The modified legend is:

"Figure 12. Analysis of the CO_2 difference between the control run (BEP_MYJ) and each of the other four sensitivity runs over two one-month periods. The colored bars show the monthly mean difference whereas the black lines indicate +/- one standard deviation of the monthly values. The results are shown for two periods of the day (afternoon 11-16 UTC, nighttime 00-05 UTC) and for three land use types (urban, crop and the others)."

The modified text is:

"We also accessed the respective contributions of anthropogenic and biogenic fluxes to the simulated spread of CO_2 concentrations using different physics schemes. This allows an estimate of the impact of uncertainties in the atmospheric transport modeling along with that of the impact of the various flux contributions. Figure 12 shows the statistics of the differences in simulated anthropogenic and biogenic CO_2 at approximately 20 m AGL between the control run (BEP_MYJ) and each of the other four sensitivity runs."

Page 11 line 13 – A strong conclusion for the use of KNN outlier removal that is not justified, see above.

Please see the answer above.

Figure 12 – Why use a cumulative distribution and not a histogram – what are the authors trying to show with this choice?

Note: Figure 12 is now ranked as Figure 13 in the revised manuscript.

Figure 13b is changed to a basic histogram as suggested. The cumulative histogram makes it easy to quantify the fraction of data below or above a threshold, while the basic histogram is much easier for a visual representation of data distribution. Given that these two formats make no difference in representing the core findings of the study, we therefore changed it to a basic one to make the content more readable.

Technical Comments

Page 8 line 18 – suggest "(the SAC station had unfortunately measurement gaps)" -> "(unfortunately the SAC station had measurement gaps)" and dates of gap added.

Text is modified as suggested:

"(unfortunately the SAC station had measurement gaps from May 3rd to June 23rd and from July 7th to July 12th)"

General note on figures – Rainbow colour schemes should be replaced with perceptually uniform colour schemes, and red-green colour schemes should be avoided due to Colour vision deficiency (colour blindness).

We thank the reviewer for the suggestion. The rainbow colour scheme in Figure 8 has been replaced with a perceptually uniform colour scheme.

Figure 4 – Both the dark blue-filled contour and red contour are called the 'threshold' but the two are not in agreement (red contour seems to be the correct one).

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

We thank the reviewer for having spotted the mistake. We have corrected the caption. The filled contour with the blue colormap is from the minimum anomaly score (dark blue) to a threshold value (light blue), so that it is the light blue and the red contour that indicates the threshold. The modified text is as follows:

"The shade of blue area indicates the anomaly score for each point, with the minimum in dark blue and the threshold value in light blue."

Figure 6 – A note to remind the reader that this figure is for January only would be helpful.

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

We have added "in January" in the caption.

Figure 7 – This figure is dense, which hinders clarity. I suggest the wind direction and mf time series to be moved to new windows with a shared x axis. The boundary layer height should also have a larger contrast to make it more visible.

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

Figure 8 is changed 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.

Figure 10 – Showing the line of transect for the south-north slice on the lat-lon plots would make interpretation of the figure clearer.

Note: Figure 10 is now ranked as Figure 11 in the revised manuscript.

Figure 11 is changed as suggested. We have added a white dash line to show this south-north transect.

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