Urban atmospheric CO2 plumes from space – Part 1: Atmospheric modeling of the urban boundary layer

The manuscript presents the first part of an evaluation of modelling CO_2 with WRF-Chem over the Paris area. The focus is on the impact of various parameterizations of boundary layer processes and urban physics. Specifically four urban canopy represenations and three PBL-schemes were evaluated against meteorological observations, surface flux measurements, CO2-monitoring and lidar measurements. The study concludes that the BEM urban scheme coupled with the MYJ PBL scheme matches the observations best, particularly during winter.

This article is a contribution for urban atmospheric transport modelling and a step to link measurements to emission inventories. The goal and design of the study are very clear. The observational dataset is impressive. The manuscript could benefit from some shortening by trying to avoid repeating information and some of the conclusions should be supported by more evidence or worded more carefully. In that regard, the model comparison should be supported by adding confidence intervals or significance testing. Overall, I believe the manuscript is valuable and suitable for publication after revision.

General Comments:

- Many paragraphs repeat a similar discussion, but for different variables. This makes the article quite long and sometimes difficult to keep track of what's happening at a given location/time, since the information is distributed over many pages. The authors should consider shortening the text. E.g. by first presenting the results and then discussing them once per location/time.
-
- There are several limitations mentioned in the text, but then mostly ignored in the discussion. E.g.:
    - different methods to determine the PBL height in observations vs models. How does this affect your comparison, is there a systematic difference reported in the literature for the two methods? **Is a quantitative comparison even meaningful**?
    - Limitation in the emission inventory. Missing emissions, wrong emission heights, how accurate are the temporal emission profiles? How does that affect your comparisons? Is it the same at all compared stations, or would you expect some of them to do worse (e.g. because there is a strong emission source nearby)
- What is the set-up for the different schemes? How were the parameters chosen?
    - surface properties (albedo, emissivity, ...)
    - How does it affect the surface/building energy balance and in turn your flux and PBL metrics?

    - How representative is the choice you made? Would a simpler scheme with better choices outperform a more complex scheme?
- Are the reported differences significant?

Specific Comments:

Title: I would reconsider the title, as there is no direct link to space in this article (only in "part 2").

L96: What do you mean by "intensive period"?

L96: Are the selected periods representative for winter/summer? Are they close to what can be expected from climatology or were there any extreme or unlikely events?

L126: Are the urban properties the default values or case-specific? (E.g. anthropogenic heating, building properties (#layers, albedo, emissivity,…), vegetation properties (height, LAI, albedo, ..), …)

L141: What's the expected impact of the nudging on meteorological and CO2 diagnostics? How does the nudging interact with the evaluation of the physics schemes? E.g. would the improvement of the e.g BEM over SLUCM be similar without nudging, better or worse?

L169: How accurate is the inventory, how much of your model/measurement-mismatch would you attribute to the inventory?

L179: What is the large-scale background CO2 based on? What is the contribution of the outer domains to your innermost domain? Does the model at upwind rural stations match the observations?

L189: Yes, but it does differ temporally too. I.e. traffic timing is not the same everywhere. Traffic only has one temporal profile, but inbound and outbound streets could behave very differently? How would you expect this to affect the CO2-comparison?

L194: This is quite a substantial simplification. While it is understandable that it was done this way, it would definitely affect how much vertical mixing is "required" from the model to match observations. And should be carfully considered comparing CO2 later on

L295: Why are the meteorological variables not interpolated to the measurement height? Especially in regions with strong gradients one would expect a better match.

L301: This seems to be the period when the model matches the observations best. However, wouldn't it be more meaningful to evaluate when CO2 concentrations are largest? Since the end goal is better modelling of CO2?

L307: You caution the reader here to interpret PBLH differences, but this is not clearly reflected in the articles analysis of differences in PBLH later on. Is a quantitative analysis even meaningful?

L320: It is not very clear throughout this section if the differences are an issue of the schemes in general, or just the parameters chosen in your simulations. It might help to have the total energy balance for all schemes, to judge where the discrepancies are coming from (wrong heat storage, Sensible/latent partitioning, long-wave exchange, albedo, …?) Maybe in the appendix, just to support your arguments.

L320: Throughout the section multiple statements are not well supported with (presented) evidence (e.g. what causes a higher heat flux). Referring to the full energy balance might help, or otherwise the wording should be more cautious.

L337: «indicating excessive heat release associated with ist parameterized building effects and enhanced roughness-driven mixing.». what supports this statement? Is it one, the other, or both? Is the total absorbed sunlight wrong and the increased HFX just a consequence? Is not enough energy stored in the ground/buildings, is already stored energy released? Where does the energy go, come from? Is the partitioning to latent heat wrong? How is HFX 200W/m2 at night?

L339: The modelled peaks seem to be 1-2h before the measured peaks, is there an explanation?

L344: «but the magnitude of the bias is smaller in absolute terms than in winter», this seems not to be true for all schemes

L348: «stored building heat release», couldn't this be domestic heating?

L352: «reflecting .. reduced heat storage», How does reduced heat storage reduce the overall HFX? Wouldn't it just redistribute HFX from day to night?

L403: «LH is weakly constrained by urban physics during winter», is this valid in general or just because the chosen period was dry?

Table3: The tables are generally clear! A slight visual guide could be helpful, e.g. highlighting the best value per statistic

Table3: «Data are averaged…», it is not just an average, it is multiple statistics over that period

L437: For the TKE etc. are resolved and subgrid contributions included?

L559: The wind discrepancies at SIRTA are not considered when discussing CO2 at SAC. (Which is either at the same location or close)

L579: "This improved performance of BEM compared to BEP likely reflects the inclusion of building thermal inertia and indoor–outdoor energy exchanges», which of the two, or both? You could check.

L680: You should clearly show that the difference in statistical measure (RMSE, MAE, Bias and Correlation) are significant between the model configurations. E.g. by adding significance tests or confidence intervals

L689: You should be more careful with the conclusions in this section. Based on the assumptions in your emissions different factors are important in explaining your observations. Lack of sources e.g. human respiration would lead to an overall underestimation. How would the simplification of a unified temporal profile across the entire domain affect local measurements? How the emission height, plumes, problem in VPRM due to wrong "color" of pixels, mismatch in wind directions. I recommend stating you observations but not concluding that one urban or PBL parameterization is strictly better. Since not enough evidence is provided.

L698: The peaks seem to not to be nocturnal, but morning rush-hour related.

L711: Differences seem not to increase over night, so is it really accumulation/trapping?

L729: “compensation effect” vs “smoothed out in the diurnal averaging”. A smoothing is not really a compensation.

L753f:”This behavior is …”, how can you exclude the possibility that nighttime emissions are wrong?

L759: “well reproduced, particularly during the morning …” the best match is not in the morning, but in the afternoon from ~10UTC onwards (12LT in summer)

L776: “During summer nights, ….”, how do we know this, what is a prolonged accumulation period?

L784: It seems to me that mid-cost stations are more “urban” on average, how do you exclude that the differences to Picarro are not a sample bias?

Fig 10: What is the hourly background concentration? There seems to be very little local contribution?

L807: “..near-surface meteorological variables… rather than acting independently”, what is that supposed to mean?

L841. Punctuation missing

L869: “.. differences among urban physics schemes consistently manifest …, rather than trough changes in large scale meteorological forcing” What does that mean, how would the urban physics scheme influence the large scale meteorological forcing?

L937 “accounting for observational uncertainty when evaluating urban CO2 simulations”, Yes, but what about the model uncertainty that is arguably bigger.

L1001ff: Unnecessary repetition of L963ff