Response to Reviewer Comments

Urban atmospheric CO_2 plumes from space – Part 1: Atmospheric modeling of the urban boundary layer

We thank the reviewer for a thorough and constructive reading of the manuscript. The comments have prompted a substantial revision of the presentation and, in several places, additional analysis. Below we reproduce each comment (in italics) followed by our response. Where a specific change to the manuscript is planned, we describe it explicitly; changes will be highlighted in the revised manuscript. We have organized our responses in the same order as the review: General comments followed by specific comments keyed to the original line numbers.

General Comments

1. Repetition and manuscript length

Comment: *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.*

Response: We agree that the current structure (one subsection per variable, each walking through JUS-winter, JUS-summer, SIRTA-winter, SIRTA-summer in turn) creates repetition across Sections 3.1–3.6. In the revision we will restructure the results section around location/season blocks rather than variable blocks: for each site and season we will present the full multi-variable picture (HFX, LH, u*, TKE, PBLH, CO_2) together, followed by a single integrated discussion of the underlying mechanism, instead of re-deriving the same mechanistic narrative separately for every variable.

2a. Limitation: differing PBLH retrieval methods

Comment: *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?*

Response: This is a fair criticism. We note this caveat at L307 but did not subsequently discuss its implications for the interpretation of our quantitative PBLH bias statistics. In the revised manuscript, we will clarify this limitation and summarize the relevant evidence from the literature on differences between bulk-Richardson-number PBLH and aerosol-backscatter-derived mixed-layer height (e.g., Kotthaus et al., 2023). This study shows that aerosol-based retrievals generally identify the top of the aerosol-mixed/residual layer, whereas Richardson-number methods identify a thermodynamically defined transition in turbulent mixing. Therefore, the two quantities are not always equivalent, and their differences are strongly dependent on boundary-layer regime.

We will therefore revise the discussion to emphasize that our PBLH comparisons should not be interpreted as a direct validation of identical physical quantities, but rather as an assessment of consistency between commonly used observational and model-based PBL height diagnostics. We will also acknowledge that part of the apparent PBLH bias may arise from methodological differences in the PBLH definitions rather than model deficiencies alone.

2b. Limitation: emission inventory uncertainty

Comment: *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)?*

Response: We agree that uncertainties in the emission inventory, including missing sources, emission heights, and temporal profiles, require a more explicit discussion. In the revised manuscript, we will add a dedicated paragraph addressing these limitations. Specifically, we will

- quantify the reported uncertainties of the inventories. Super et al. (2020) showed that while uncertainties in total domain emissions are relatively small, uncertainties increase substantially after spatial and temporal disaggregation, reaching several tens of percent at the grid-cell scale for CO_2. These uncertainties are sector and location-dependent, with point sources and spatial allocation of emissions representing important sources of uncertainty.
- discuss the missing 8–10% respiration contribution already noted at L182 and its potential impact on simulated CO_2 concentrations.
- clarify how inventory uncertainties may affect the comparison with observations.

Although the same emission inventory is used in all model configurations, inventory-related errors are expected to be spatially heterogeneous. Stations located close to strong local sources, such as major traffic corridors (e.g., the Boulevard périphérique) or point sources, are more sensitive to uncertainties in emission magnitude, temporal profiles, and effective emission heights. Conversely, stations with fewer nearby sources are expected to be more representative of the larger-scale urban background and less affected by local inventory uncertainties. Therefore, part of the station-to-station variability in model–observation differences may reflect uncertainties in the emission representation rather than differences in urban physics alone. However, because all simulations use identical emission inputs, these uncertainties do not affect the relative comparison among urban physics configurations.

3. Scheme configuration and parameter choices

Comment: *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?*

Response: All four urban configurations use the WRF default surface parameter tables (albedo, emissivity, roughness length, roof/wall/road fractions, thermal properties, anthropogenic heat profiles) for the corresponding urban categories, with building morphology (height distribution, plan/frontal area index) supplied by the 10 LCZ-resolved urban classes of the Demuzere et al. (2023) 100 m hybrid land-cover product rather than a single uniform urban type. No manual tuning of any scheme's internal parameters was performed; our intent was to evaluate the schemes "out of the box" as they would typically be used by the wider GHG-modeling community. We will add this clarification explicitly to Section 2.1. We agree with the reviewer that scheme ranking can in principle be sensitive to parameter choice, and that a well-tuned simpler scheme could outperform a poorly-configured complex one; we will add this explicitly as a caveat to our conclusions and note it as a natural direction for follow-up sensitivity work (e.g., varying anthropogenic heat release and thermal admittance within SLUCM, etc).

4. Statistical significance of reported differences

Comment: *Are the reported differences significant?*

Response: The largest contrasts highlighted in the manuscript, particularly between the single-layer schemes (No_URB, SLUCM) and the multi-layer schemes (BEP, BEM), are substantially larger than the expected observational uncertainties and are therefore unlikely to result from sampling noise. For example, the winter HFX correlation increases from $R \approx 0.36$–0.37 for No_URB/SLUCM to $R = 0.64$ for BEP/BEM, and the winter PBLH bias changes from strongly negative (≈ −140 m) to near-zero or positive values. However, we agree that some finer-scale comparisons are not sufficiently robust to support a strict ranking of schemes. In particular, the differences between BEP and BEM for TKE and winter CO_2, and most summer PBLH and CO_2 differences, are small relative to the observational and retrieval uncertainties.

Specific Comments

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

Response: Thank you for this helpful comment. We agree that the phrase "from space" in the title may be misleading, as Part 1 focuses on the atmospheric modeling of urban CO_2 plumes and on sensitivity tests of the PBL and urban physics schemes in WRF.

To better reflect the scope of Part 1, we have therefore removed "from space" from the title. The revised title is:

"Urban atmospheric CO_2 plumes – Part 1: Atmospheric modeling of the urban boundary layer"

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

Response: We use "intensive period" in the sense of a field-campaign Intensive Observation Period: a window during which the full multi-instrument observational network (Doppler lidars, ceilometers, flux towers, CO_2 sensors) was operated concurrently and continuously. We will replace "intensive period" with "intensive observation period".

Comment: *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?*

Response: The selected periods were chosen as representative case studies for winter and summer conditions based on the availability of comprehensive observations required for model evaluation and LES simulations (part 2 of the paper), rather than to provide a climatological assessment. The winter period (10–31 January 2024) and summer period (1–20 June 2024) correspond to contrasting seasonal boundary-layer regimes and include continuous measurements from the full observational network used to constrain and evaluate the simulations.

We acknowledge that the winter period includes a pronounced anticyclonic, low-wind episode around 16 January associated with enhanced CO_2 accumulation. This event represents an extreme but physically relevant urban boundary-layer situation and clarify that it is interpreted as a high-impact case rather than as a climatological average winter condition. The selected periods therefore provide realistic test cases for assessing the sensitivity of urban schemes and PBL processes, while not being intended to represent the full seasonal climatological variability.

Comment: *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, ...))*

Response: As detailed in our response to General Comment 3, these are WRF default parameter tables applied to the LCZ classes from Demuzere et al. (2023); they are not manually tuned for the Paris domain.

Comment: *L141: What's the expected impact of the nudging on meteorological and CO_2 diagnostics? How does the nudging interact with the evaluation of the physics schemes? E.g. would the improvement of e.g. BEM over SLUCM be similar without nudging, better or worse?*

Response: The FDDA configuration used here nudges the three-dimensional wind, temperature and moisture fields toward observations (the standard WRF FDDA implementation), and does not nudge the CO_2 tracer fields at all. Because nudging chiefly constrains the synoptic-scale state common to all four urban configurations, we expect its direct effect on the relative differences between urban schemes to be small. We cannot fully rule out an indirect effect (e.g., nudging-driven differences in advected background CO_2 or wind direction feeding back on local turbulence) and a dedicated un-nudged sensitivity run is beyond the scope of the current revision.

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

Response: See our response to General comment 2b. We will add the reported inventory uncertainty ranges from Super et al. (2020) for TNO and note that Airparif does not publish a formal uncertainty budget; we will explicitly state that we cannot cleanly partition model–observation mismatch between emission-inventory error and transport error with the present set of experiments and will avoid attributing residual bias solely to boundary-layer physics.

Comment: *L179: What is the large-scale background CO_2 based on? What is the contribution of the outer domains to your innermost domain? Does the model at upwind rural stations match the observations?*

Response: The background/lateral boundary CO_2 fields come from the CAMS global CO_2 reanalysis (1°×1°, 3-hourly; Agustí-Panareda et al., 2023), applied at the D01 boundary and propagated inward through the one-way nested domains (D01→D02→D03) without feedback to the coarser grid, as stated in Section 2.1. Therefore, the contribution of the outer domains to the innermost domain is through the lateral boundary conditions and the transport of the large-scale CO_2 background, while the small-scale urban enhancements are generated by the higher-resolution emission fields within the inner domains. We agree that the representation of the background CO_2 field is an important source of uncertainty and will clarify this point in the revised manuscript. The agreement between modeled and observed CO_2 at rural/upwind stations depends on the accuracy of both the CAMS background and the representation of regional transport, as these sites are less influenced by local urban emissions. We will therefore discuss the rural background representation as a limitation of the model–observation comparison, noting that any bias in the imported background affects all urban simulations similarly because the same boundary conditions are used for all urban physics configurations.

Comment: *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 CO_2-comparison?*

Response: We agree this is a real simplification: the sector-specific temporal profiles used to disaggregate annual emissions are applied uniformly across all road segments of a given sector, without distinguishing inbound/outbound or radial/orbital traffic axes that can have distinct rush-hour signatures. We expect this to introduce phase errors of order an hour at stations dominated by a single nearby road with a strongly directional flow (e.g., stations close to the périphérique or major radial routes), while having less impact at stations further from dominant point/linear sources. We will add this caveat explicitly at L189 and note it as a candidate explanation for some of the timing offsets discussed at L339 and L759.

Comment: *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 carefully considered comparing CO_2 later on.*

Response: This is directly relevant to interpreting Section 3.6: because emissions are injected into the ~10 m surface layer rather than at their true release heights (e.g., rooftop HVAC, vehicle exhaust, elevated stacks), while CO_2 is measured at 33–34 m (JUS/EIF) and 60 m (SAC), schemes that under-mix will tend to trap CO_2 close to the surface and may appear to under rather than over-estimate concentrations at instrument height, or vice-versa depending on the vertical gradient. We will add explicit discussion of this effect where we compare No_URB/SLUCM versus BEP/BEM performance in Section 3.6, noting that some of the apparent skill of weakly-mixing schemes could in part reflect a compensating error between missing emission height and insufficient mixing rather than either factor considered alone. This reinforces the motivation, already noted at L1005, for future sensitivity experiments using explicit multi-level or building-resolved emission injection.

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

Response: In the present analysis, we selected the model level closest to each instrument measurement height rather than applying vertical interpolation. The lowest atmospheric model level is located at approximately 10 m, which is close to the height of the meteorological observations used in this study. Therefore, the vertical distance between the observation height and the selected model level is generally small, and we expect the impact of interpolation on the reported statistics to be limited. Nevertheless, we acknowledge that interpolation can be advantageous in situations with strong vertical gradients, and we will clarify this methodological choice in the revised manuscript.

Comment: *L301: This seems to be the period when the model matches the observations best. However, wouldn't it be more meaningful to evaluate when CO_2 concentrations are largest? Since the end goal is better modelling of CO_2?*

Response: The selected daytime windows (10–16 UTC in winter and 09–18 UTC in summer) were defined for the evaluation of surface fluxes (HFX, LH, u*, and TKE) and PBLH, because these variables are most physically meaningful during periods of developed turbulence and active surface–atmosphere coupling. They were not selected based on periods of better model performance.

We agree that CO_2 accumulation is often strongest during nocturnal and morning periods, when reduced mixing leads to enhanced near-surface concentrations. To account for this, the full diurnal evolution of CO_2 (including nighttime accumulation and morning transition periods) is shown through the diurnal cycle analysis. The statistical comparison of CO_2 concentrations follows the same daytime window used for the process-based evaluation, allowing a consistent assessment of the impact of urban physics on coupled surface–boundary-layer processes.

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

Response: See our response to general comment 2a. We stated that quantitative PBLH comparisons should be interpreted with caution because the observational and model-based PBLH diagnostics do not necessarily represent exactly the same physical quantity. However, we consider the comparison meaningful for identifying systematic differences between model configurations, particularly when interpreted together with other boundary-layer variables and the diurnal evolution of PBLH. In the revised manuscript, we will clarify that the reported PBLH biases should not be interpreted as a direct measure of absolute model error, since part of the discrepancy may arise from methodological differences in PBLH definitions. Instead, the quantitative results are used to assess the consistency and relative behavior of the different urban physics configurations.

Comment: *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 (second): 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.*

Response: The comparison presented here focuses on the overall behavior of the different WRF urban physics configurations (No_URB, SLUCM, BEP, and BEM), rather than on isolating the impact of individual parameter choices within each scheme. The differences therefore reflect the combined effects of the different urban canopy formulations and their associated parameterizations in WRF.

We agree that a complete energy balance analysis would be required to quantitatively separate the contribution of individual processes (e.g., heat storage, radiative exchange, and sensible/latent heat partitioning, etc) to the simulated differences. Such an analysis is beyond the scope of this study, which aims to evaluate the consequences of different urban physics configurations for boundary-layer dynamics and CO_2 simulations. The interpretation presented in this section is based on the available diagnostics (HFX, LH, u*, TKE, and PBLH), which provide insight into the resulting surface–boundary-layer interactions. We will clarify in the revised manuscript that these diagnostics describe the integrated response of each urban physics configuration rather than a direct attribution to individual energy balance components.

Comment: *L337: 'indicating excessive heat release associated with its 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 200 W m-2 at night?*

Response: We agree that the original wording over-attributed the causes of the BEP HFX overestimation. The available diagnostics in this study demonstrate the resulting differences in sensible heat exchange and boundary-layer response, but they do not allow us to unambiguously separate the contribution of individual energy-balance terms (e.g., heat storage, radiative exchange, anthropogenic heat release, or sensible/latent heat partitioning, albedo, etc). We will therefore revise this statement to avoid attributing the excessive HFX to a specific mechanism. Instead, we will describe it as an enhanced sensible heat release from the BEP configuration associated with its more complex urban representation, and discuss possible contributing mechanisms based on previous studies of BEP behavior (e.g., Martilli et al., 2002; Salamanca and Martilli, 2010), while acknowledging that the relative contribution of each process cannot be diagnosed here.

Regarding the large nighttime HFX values in BEP/BEM at JUS, we agree that this behavior requires clarification. Such values are not expected from direct solar forcing and may result from the representation of delayed urban heat release from building and surface thermal storage, together with prescribed anthropogenic heat fluxes. We will clarify this interpretation and acknowledge that a complete attribution would require additional energy-budget diagnostics beyond the scope of this study.

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

Response: The 1–2 h phase difference between modeled and observed HFX peaks may result from differences in the representation of the surface energy balance and its response to incoming radiation. Candidate explanations include scheme-dependent treatment of urban heat storage and the timing of heat release, differences in the partitioning between storage and turbulent fluxes, and possible effects of FDDA nudging on the modeled boundary-layer evolution. However, the available diagnostics do not allow us to isolate a single cause of this phase offset.

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

Response: Thank you for pointing this out. We agree that this statement was not accurate as a general description of the results. We will correct this sentence in the revised manuscript.

Comment: *L348: 'stored building heat release', couldn't this be domestic heating?*

Response: Yes — domestic/anthropogenic heating is a plausible and likely joint contributor to elevated wintertime nighttime HFX in the multi-layer schemes, which explicitly represent an anthropogenic heat term, whereas stored solar heat release from thermal mass is a separate, purely physical mechanism present in both BEP and BEM.

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

Response: We agree that the original wording was not precise. Our intended point was that differences in urban characteristics and surface properties between sites can modify the partitioning of available energy between sensible and other heat flux components, rather than that reduced heat storage directly reduces HFX. We acknowledge that heat storage primarily affects the timing of heat release and can redistribute sensible heat flux between daytime and nighttime periods. We will revise this sentence to avoid attributing the HFX difference to a direct reduction in heat storage and instead describe it in terms of differences in surface energy partitioning associated with the site characteristics.

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

Response: We agree that this statement was too general. The present analysis is based on a single three-week winter period, and we cannot infer a general property of wintertime urban LH from this case alone. We will revise the text to explicitly scope this conclusion to the relatively dry conditions sampled during the winter period studied. Under such conditions, limited moisture availability may reduce the sensitivity of LH to differences among urban physics configurations; however, the response may differ under wetter conditions or during periods with stronger evaporative demand.

Comment: *Table 3: The tables are generally clear! A slight visual guide could be helpful, e.g. highlighting the best value per statistic.*

Response: Thank you, we will add shading to highlight the best-performing (lowest |MBE|/MAE/RMSE, highest R) scheme in each column of Tables 3–6.

Comment: *Table 3: 'Data are averaged...', it is not just an average, it is multiple statistics over that period.*

Response: Correct, we will revise the caption to read 'Statistics are computed over the active daytime period...' rather than 'Data are averaged over the active daytime period...'.

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

Response: The modeled TKE used in this comparison is the PBL-scheme diagnostic variable, which represents the parameterized (subgrid) turbulence at the grid scale; at 900 m horizontal resolution essentially all boundary-layer turbulence is subgrid and thus captured by this term (explicit/resolved turbulence is negligible at this resolution). We will state this explicitly at L437, and add the caveat that eddy-covariance towers measure resolved, high-frequency turbulence at a point, which is not perfectly comparable to a grid-box-averaged subgrid diagnostic; this scale mismatch is a further source of the systematic TKE underestimation discussed in Section 3.2 and will be noted explicitly.

Comment: *L559: The wind discrepancies at SIRTA are not considered when discussing CO_2 at SAC. (Which is either at the same location or close).*

Response: SAC and SIRTA are indeed co-located (Table C1/E1 show near-identical coordinates). We agree that the wind-direction discrepancy identified at SIRTA should also be considered when interpreting the CO_2 comparison at SAC, since errors in the simulated flow direction can influence the transport and dilution of urban CO_2 enhancements. We will add an explicit cross-reference in Section 3.6 to acknowledge that part of the model–observation discrepancy in CO_2 may arise from transport uncertainties, in addition to differences in local mixing associated with the urban physics schemes.

Comment: *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.*

Response: We agree that the current wording attributes the improved BEM performance to specific mechanisms that cannot be separated with the diagnostics available in this study. The improvement likely results from the combined effect of the additional processes represented in BEM compared with BEP, including building thermal inertia and indoor–outdoor energy exchanges, but we cannot determine the relative contribution of each component without a dedicated energy-balance analysis.

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

Response: See also our response to general comment 4 on the significance of the reported differences. We agree that RMSE, MAE, MBE, and correlation coefficients alone do not provide a formal significance test. However, our interpretation is based on the magnitude and consistency of the differences across multiple diagnostics. The largest contrasts between single-layer (No_URB, SLUCM) and multi-layer (BEP, BEM) schemes are much larger than expected observational uncertainties and are unlikely to result from sampling variability, while smaller differences between individual schemes are interpreted more cautiously. Since all configurations are evaluated against the same observations and periods, the comparisons are paired and are intended to assess relative scheme behavior rather than perform hypothesis testing.

Comment: *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 your observations but not concluding that one urban or PBL parameterization is strictly better. Since not enough evidence is provided.*

Response: We thank the reviewer for this important comment. We agree that uncertainties beyond the urban physics representation (e.g., missing respiration sources, emission temporal profiles, emission heights, VPRM-related uncertainties, and wind-direction biases affecting CO_2 transport) affect the interpretation of model–observation differences. Since all simulations use the same emissions and boundary conditions, these factors do not explain the differences between urban physics configurations directly, but they do limit the interpretation of absolute model performance.

We will therefore revise the discussion and conclusions to avoid implying that one urban/PBL scheme is universally superior. Instead, we will state that a given configuration provides better agreement with observations under the emissions and boundary conditions used in this study, while acknowledging that separating urban physics effects from emission and transport uncertainties requires additional sensitivity experiments beyond the scope of the present work.

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

Response: We thank the reviewer for pointing this out. We agree that the CO_2 peaks are primarily associated with the morning rush-hour emission enhancement, occurring on top of elevated nighttime concentrations maintained by reduced nocturnal mixing. Our previous wording overemphasized nocturnal accumulation as the direct cause of the peak. We will revise the text to describe this process as the combined effect of morning traffic emissions and the residual influence of overnight stable conditions, rather than as a purely nocturnal trapping event.

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

Response: We agree that “accumulation/trapping” is too strong given that the differences do not progressively increase overnight. We will revise the wording to refer instead to enhanced nighttime CO_2 retention and reduced dilution under weak mixing conditions, without implying continuous

Comment: *L729: 'compensation effect' vs 'smoothed out in the diurnal averaging'. A smoothing is not really a compensation.*

Response: Agreed, these are two distinct concepts, and we conflated them. We will remove “compensation effect” and describe this as an averaging/smoothing effect: episodic nighttime overestimation is attenuated in the mean diurnal cycle, making No_URB/SLUCM appear to better reproduce the observed amplitude despite poorer agreement during individual events.

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

Response: We cannot fully exclude errors in the nighttime emission with the current design. We will therefore revise the text to present the representation of nocturnal mixing/stability and potential errors in the nighttime emission temporal profile as two plausible, non-exclusive explanations for the summer nighttime CO_2 underestimation, rather than attributing it solely to boundary-layer physics.

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

Response: We thank the reviewer for pointing this out. We agree that the wording “particularly during the morning transition” was too specific. The best agreement occurs from approximately 10 UTC onward in winter and from approximately 08 UTC onward in summer, extending through the afternoon. We will revise the sentence to reflect this broader period of good agreement rather than specifically attributing it to the morning transition. “During daytime, simulated CO_2 mixing ratios are generally well reproduced, particularly from approximately 10 UTC onward in winter and 08 UTC onward in summer, when boundary-layer mixing is well established.”

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

Response: We agree that "prolonged accumulation periods" was not sufficiently defined and implies a specific accumulation mechanism that is not directly demonstrated by our analysis. We will remove this wording and simply state that the nighttime CO_2 underestimation is most pronounced during periods of elevated observed concentrations.

"During summer nights, all configurations tend to underestimate CO_2 peaks, with the largest discrepancies occurring during periods of elevated observed concentrations. This behavior may be related to differences in the representation of nocturnal boundary-layer mixing, as the simulated turbulence levels and near-surface exchange may not fully reproduce the observed nighttime conditions. The persistence of non-zero friction velocity and TKE in the simulations suggests that some mechanical mixing remains active, which may limit near-surface CO_2 enhancement compared with observations. However, uncertainties in nighttime emission representation may also contribute to these differences. Overall, the results highlight a common limitation of the tested urban physics configurations in reproducing summer nocturnal CO_2 variability."

Comment: *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?*

Response: We agree that differences in station characteristics may contribute to the different levels of model–observation agreement between the Picarro and Mid-cost networks. While differences in measurement accuracy and noise characteristics likely play a role, we cannot fully exclude a sampling effect related to station siting, urban intensity, and exposure to local emission sources. We will revise the text to acknowledge that the observed differences may result from a combination of instrument characteristics and network representativeness.

"For both seasons, model–observation agreement is generally higher for the Picarro network compared to the Mid-cost sensors. This may reflect differences in measurement accuracy and stability between the instrument types, but could also partly result from differences in station characteristics, including urban exposure, local sources, and atmospheric representativeness."

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

Response: We thank the reviewer for this comment. The concentrations shown in Figure 10 represent the total near-surface CO_2 mixing ratio, including the large-scale background contribution and the local enhancement from regional and urban emissions. The background contribution is indeed a substantial component of the total signal, but the local urban enhancement is still significant, particularly during winter pollution events and at urban sites. The objective of Figure 10 is to evaluate the ability of the different urban physics configurations to reproduce the total observed variability, including the timing and magnitude of CO_2 enhancements. We do not show the hourly background concentration in this figure because it would not directly help evaluate differences among urban physics schemes, which are assessed using the same background forcing in all simulations.

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

Response: We agree that this sentence was unclear. Our intended meaning was that CO_2 variability is controlled by the combined influence of boundary-layer processes (surface fluxes, turbulence, and PBLH), rather than by any single near-surface meteorological variable alone. We will revise the wording to make this coupling clearer and remove the ambiguous phrasing.

Comment: *L841: Punctuation missing.*

Response: Thank you, this will be corrected.

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

Response: We agree that this sentence was confusing. The urban schemes do not modify the imposed large-scale meteorological forcing, which is identical across all simulations. The intended meaning was that the differences among simulations arise from the different representations of urban surface processes and their effects on surface energy partitioning, turbulence, and boundary-layer development. We will revise the sentence to make this distinction clear.

“This study demonstrates that the representation of urban surface processes exerts a first-order control on simulated planetary boundary layer (PBL) dynamics and, consequently, on near-surface CO_2 mixing ratios over Paris. Across both winter and summer periods, differences among urban physics schemes arise from their representation of surface energy partitioning, turbulence generation, and boundary-layer development under the same large-scale meteorological forcing.”

Comment: *L937: 'accounting for observational uncertainty when evaluating urban CO2 simulations', Yes, but what about the model uncertainty that is arguably bigger?*

Response: We agree that observational uncertainty is only one component of the overall uncertainty in urban CO_2 simulations. The results of this study also demonstrate that differences among urban and PBL parameterizations can introduce substantial variability in simulated CO_2, particularly under winter conditions. We will revise the sentence to place observational uncertainty in the broader context of both measurement and model structural uncertainties, rather than implying that observational uncertainty is the dominant limitation.

“The consistently better agreement with the Picarro network compared to the Mid-cost network across both seasons suggests that measurement characteristics, including instrument precision, calibration stability, and response time, contribute to model–observation differences. However, observational uncertainty represents only one component of the overall uncertainty in urban CO_2 simulations. The differences among urban and PBL parameterizations evaluated in this study introduce additional variability in simulated CO_2, particularly under winter conditions when boundary-layer processes strongly influence tracer accumulation and dispersion. Therefore, both measurement uncertainty and model structural uncertainty should be considered when interpreting model–observation comparisons, especially when differences among configurations are relatively small, as in summer.”

Comment: *L1001ff: Unnecessary repetition of L963ff.*

Response: We agree that the two paragraphs repeat the motivation for Part 2. We will remove the redundant paragraph in the Conclusions that restates material already presented in the Discussion (around L963).