We thank Reviewer 2 for the very constructive suggestions for improvement, that helped us extending the manuscript and improving the simulation results significantly. We will first give a short general summary of the changes. In the following, we will answer the points brought up by Reviewer 2 in detail.

Summary of the Changes

We decided to alter the simulation setup and extend the ICON-ART simulations. This decision was based on the question of the validity of the background assumptions. This aspect was brought up by both reviewers. Especially the question if a larger domain size and longer transportation time are necessary could not be answered sufficiently in the framework of the previous simulation setup. The new setup includes a second outer domain, covering a larger area. A longer transportation time was simulated. We describe the new setup in the revised manuscript.

The new setup increased the general agreement between the prior simulation results and the observed XCO2'.
At the same time, the estimated emissions of Thessaloniki slightly decreased.
The statement of an underestimation of ODIAC's emissions is still valid, especially when considering actual ODIAC data for 2021 and 2022, which became available in the meantime in ODIACv2024 dataset.
To answer the question of the effect of out-of-domain sources the former reverse approach to investigate far-field emissions was replaced by separating the far-field background and longer transportation times and by redoing the emission estimation in different configurations. Both effects affected the resulting emission estimates only minimally.

As was suggested by Reviewer 1, we further investigated the biogenic contribution of the variability, which showed that the biogenic sink plays an important role during the summer month in 2022 but is of minor importance for October 2021. This brought us to the idea of separating the whole observation period in monthly sub-periods instead of selecting days with a better prior agreement.
This separation showed that the agreement was superior for October 2021. Concurrently the largest underestimation of ODIAC's emissions could be seen in this month. The underestimation was less pronounced for the summer.

In summary these changes significantly increased the agreement between observed and simulated XCO2. The extended simulation shows that a larger domain or an increase in simulation times have little impact on the resulting emission estimates. Finally, the results could be refined temporarily indicating a stronger influence of the biogenic sink for the summer month. The underestimation of the emissions was found to be most pronounced in October 2021, where also the best agreement between simulation and observation is visible.

Specific Replies (Reviewer Comments in blue)

Line4&5: Please spell out the full names of 'EDGAR' and 'ODIAC' upon first mention in the abstract for clarity, especially for readers who may not be familiar with these acronyms.

We added the acronyms in the text.

Line 118: The reported XCO2 enhancements reached up to 2.03 ppm. I agree with the other reviewer that this signal appears relatively modest. Could the authors clarify the distance between the two instruments, as well as the wind conditions on that particular day? It is possible that the relatively small enhancement was influenced by the instrument spacing, mild wind speeds, or variable wind directions.

We added the wind conditions for the given day, which showed lower-than-usual wind speeds. We expect the signal to be smaller than in previous campaigns focusing on very big cities, as Thessaloniki is reported to have significantly lower emissions. Thereby, Thessaloniki is a more difficult target for the quantification of its emissions. We added a comparison in emission strengths to a previous study targeting LA to the introduction.

It would be helpful to compare this result with findings from similar urban campaigns. For example, in the early Berlin study (doi:10.5194/amt-8-3059-2015), a moderate XCO2 enhancement of approximately 4 ppm was observed at the downwind site.

The gradient is compared to observed gradients from the Tokyo Campaign where gradients up to 9.5 ppm were observed.

Line 124: The authors mentioned that simple models, such as the box model, are not suitable for this study and instead proposed a more precise simulation using ICON-ART. Could the authors clarify whether ICON-ART has previously been applied to greenhouse gas or trace gas emission studies? If so, referencing relevant prior work would help support its use in this context.

We added two references where ICON-ART is used for inversion applications.

Line 141: the authors state that they "set the surrounding background concentration to 0 ppmv". Could the authors provide more justification for this assumption? What are the characteristics of the surrounding areas, and is there evidence to suggest they have no significant influence on the target region? Including a spatial distribution of XCO_2 derived from satellite observations (or from CAMS XCO2 predictions?) could help assess whether notable emission sources are present in the surrounding areas and validate this assumption.

For a better justification of the assumptions concerning the advected background, we extended the simulation by a broader domain where more emission sources contributed to the background that is present in the simulation. We could show that the emission estimates did not differ significantly between explicitly taking this far-field background into account or considering only the emissions of Domain 1.

Line 199: The purpose of the spatial regridding is not entirely clear for me. Specifically, it's unclear why four 1 km × 1 km pixels (from the original ODIAC inventory, which already matches the simulation resolution) are merged into larger 4 km² pixels for the city with the highest emissions. Since the ODIAC inventory has a comparable resolution to the model grid, why not retain the original 1 km × 1 km pixel resolution for emission allocation? Please clarify the rationale behind this regridding step and explain how it supports the goal of rescaling the inventory after the simulation runtime.

The ICON grid is an icosahedral grid, while the other data sources are provided in latitude-longitude gridding. Therefore, a regridding is necessary. To make this more clear, we moved the image of the ODIAC inventory regridded to the Icosahedral ICON grid as used in the simulation from Figure 10 to an earlier position.

Line 214: The sentences of "Outside the subdivided area the emissions of the ODIAC inventory remain unchanged. The scaling factors, were restricted so that the prior could only be scaled in the range (1/6, 6), that was empirically chosen." are unclear to me. Could the authors clarify why emissions outside the subdivided area were kept constant? Additionally, why restricting the scaling factors to the (1/6, 6) range, and how was this empirical range determined?

To clarify the chosen approach, we extended the paragraph.

Line 224: the authors use two days of closely spaced measurements (~500 m apart) to assess the impact of spatial heterogeneity. Based on Figure 2, I assume the first instrument was located at Campus and the second at Seych Sou, which appears to be a more mountainous region. It would be helpful to explicitly state the locations of these measurements in the manuscript to provide better context for the reader.

Both sites were located on campus. To improve clarity, we extended Figure 2 to include a detailed view of the campus.

Additionally, I'm curious about the wind conditions during these two days. Was there any prevailing wind direction that could have transported emissions from one site to the other? If so, this might have influenced the observed enhancements and should be discussed as a potential factor in the interpretation. A discussion of wind-related differences between the sites could further strengthen the assessment of spatial heterogeneity and its impact on the measurements.

The wind conditions for the days were added in the text. Also the complexity of the wind fields was illustrated exemplarily in Appendix B.

Furthermore, during the calibration days, there appear to be significant biases of up to 1 ppm. Any reasons for this (e.g., could they be related to higher solar zenith angles which should be filtered?). This level of bias is comparable to the ΔXCO2 signal in this study used to estimate emissions.

The ΔXCO2 shows high discrepancies for individual pixels for both configurations, the 500 m distance measurements, and the side-by-side observations.
As proposed by Reviewer 2, we checked the dependency of the solar zenith angle (see Figure A).

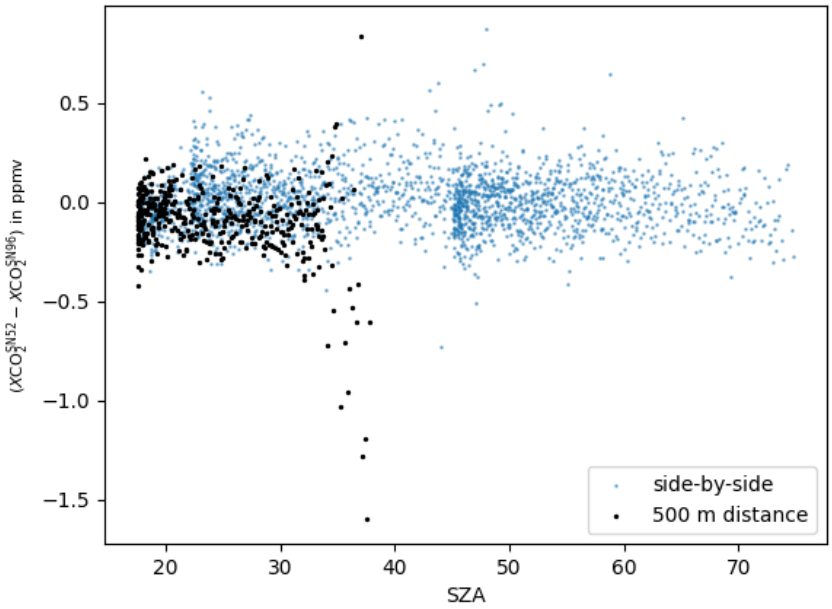

Figure A: Comparison of the SZA dependence of the side-by-side and 500m-distance observations.

For the 500m-distance observations, we find that the highest discrepancies are visible for the highest observed SZAs. however, the absolute SZA for these instances are still lower than 40 degree. As there are not many observations, it could also be that, e.g. the sun was not perfectly centered for a period at the start/ end of the measurement period where the SZA was minimal on that given date. The outliers are similarly pronounced for the side-by-side observations, e.g. in Figure 6 of the Preprint around 418 ppm. The impact of the outliers is much smaller when binning into 10-minutely bins.

The no scaling XCO_2' values appear to be significantly lower in October compared to June and July, which does not align with the observations. Could this higher discrepancy in October relate to an underestimation of biogenic sources in the ICON-ART model? Or could this discrepancy be related to the background removal method? Specifically, the use of a constant daily background based on the 5th percentile may suppress meaningful variability and potentially remove part of the signal of interest. Given that two EM27/SUN instruments were deployed, was it considered to use the upwind site as a dynamic background reference? Alternatively, if satellite data are available, would using upwind satellite observations provide a more representative background than the current approach?

The scaling factors are applied to the full time-series, so they are the same for October and June, however with the overall scaling factor the simulated enhancement based on the inventory is visibly lower in October than the observations. We therefore replaced the sub-sample investigation by individually optimizing the months, which indicates a strong annual variability of the discrepancies between inventory and measurements.

We implemented the technical comments raised by the reviewer.