Reply to reviewer #1

We would like to thank the reviewer for the critical reading and valuable suggestions. In the following, we address the points one by one presenting our replies in dark blue and changes to the manuscript in dark red.

This is an interesting and impactful paper. I recommend publication after attention to the comments below.

Major Comments

- The rationale for referencing prior work seems somewhat arbitrary. There are notable recent efforts with similar observations and similar modeling that are omitted (although not combined). A more thorough discussion of the recent literature would help the reader put this work in appropriate context.

The references covered prominent examples of urban CO_2 emission estimation studies, but we agree that recent studies employing similar observation and modeling approaches were missing. We added the following paragraph to the introduction section:

Networks of low- to mid-cost CO_2 sensors have already been deployed in other cities. A prominent example is the Berkeley Atmospheric CO_2 Observation Network (BEACO2N), a dense network of 35 nodes of CO_2 and air pollution sensors in the San Francisco Bay area (Shusterman et al., 2016), which was recently extended to other cities including Los Angeles (Kim et al., 2025), Glasgow, Providence and Heidelberg. Other examples are the Beijing–Tianjin–Hebei (JJJ) carbon monitoring system with 134 CO_2 sensor sites (Han et al., 2024) and a mid-cost CO_2 sensor network with eight sites in the city of Paris (Lian et al., 2024). In an inverse modeling study utilizing a mesoscale Lagrangian particle dispersion model, Turner et al. (2020) showed that the BEACO2N network successfully captured the reduction of CO_2 emissions during the COVID-19 pandemic.

Furthermore, we extended the sentence referencing GRAMM/GRAL studies at urban scale as follows:

Here, we describe one of these model systems, GRAMM/GRAL, which was previously used for air pollution simulations in the city of Zurich (Berchet 2017a,b) *and for optimizing the design of a CO_2 measurement network in the city of Heidelberg using an observing system simulation experiment (OSSE) (Vardag and Maiwald, 2024)*.

Furthermore, in response to a comment by reviewer #2 we added more information on the capability and limitations of a model like GRAMM/GRAL operating at 10 m resolution, where we cited additional urban modeling studies including CFD approaches (see our response to reviewer #2).

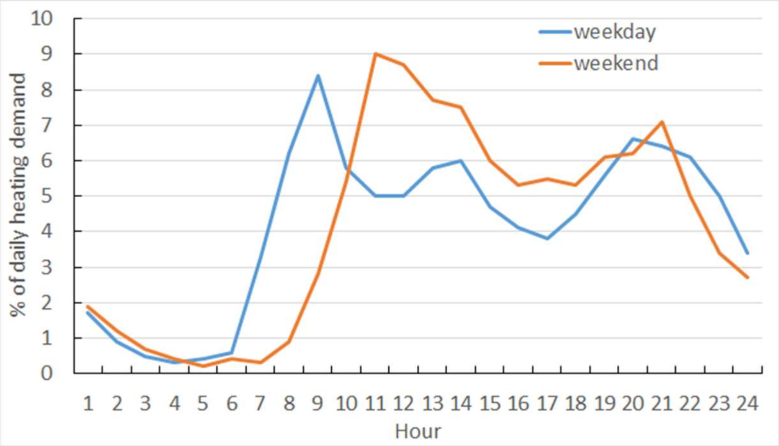
- It would be helpful to also include the spatial and temporal resolution of each input dataset in Table 1.

We added this information to Table 1. The resolution/accuracy was between 3 cm (land use cadastre of the city, vector data set) and 100 m (CORINE land cover). Regarding temporal resolution we added the following sentence:

Most data sets are static with update cycles between one and five years. When available, the reference year is included in the description. The two Sentinel-2 satellites together provide global coverage in five days. Europe is covered more frequently, but clouds lead to irregular sampling and corresponding data gaps of up to a few weeks, especially in winter.

- Why are there 3 peaks in heating/hot water demand diurnal profile? I understand this is based on a simulation from CESAR-P, but what is the behavioral reason, especially for the 3-6 am peak and then the 9-12 peak? Homes then offices? Surprising to me that it's not smoothed more in the morning

We agree that the profile didn't look very realistic. The time profile was based on a heating demand profile extracted from the CESAR-P simulation software. We had a closer look at the original publication of the Swiss Society of Engineers and Architects (SIA norm 385/1) and found much more realistic profiles of hot water demand for residential buildings for typical weekdays and weekend days (see Figure below). We implemented this new data set and regenerated all figures affected by the changed profile. The changed profile led to a small improvement in the representation of the seasonal mean diurnal CO_2 cycles in Figure 15 but had an almost negligible impact on Figures 13 and 14, which are based on afternoon mean concentrations.



- Additional quantification of the uncertainties in each aspect of the model would be helpful.

This is very challenging task (for any transport model) and in our view well beyond the scope of the present study. There is extensive literature on evaluation of the GRAMM/GRAL model system, which we are now citing in Section 2.2. as follows:

The performance of GRAMM/GAL in terms of representing meso- and microscale flows and pollutant dispersion has been evaluated extensively in previous studies. Using a similar setup as in our study, May et al. (2024) investigated how well the flow in the complex topography of the city of Heidelberg, Germany, is represented by the model. They found very good agreement at 11 out of 15 measurement sites according to the performance criteria for mesoscale air quality models formulated by the European Environment Agency (EEA, 2011). At all 15 sites, the performance was within the extended benchmark criteria proposed by Oettl and Verati (2021) for the more challenging conditions in complex terrain. Furthermore, Oettl and Verati (2021) showed that winds simulated by GRAMM in Alpine topography are at least as accurate as those simulated by the numerical weather prediction model WRF at comparable resolution. The high quality of tracer dispersion simulated by the GRAL model has been demonstrated in several studies comparing model results with wind tunnel and tracer release experiments (Oettl, 2015) and with air pollutant measurements in street canyons (Oettl and Uhrner, 2011) and across a whole city (Berchet et al., 2017b).

It is also important to note that whenever a new version of GRAL is released, the model is evaluated against a large number of validation data sets (e.g. tracer release experiments) to demonstrate compliance with Austrian, German and other national guidelines for dispersion models. These tests are extensively described in the GRAL documentation available at https://gral.tugraz.at/download/documentations/

Furthermore, we added the following sentences at the end of the results section summarizing potential sources of model error:

Future studies should investigate in more detail the main sources of model errors, which include: (i) Errors in representing the mean flow, especially wind speed, and its variability across the city. Our results suggest that these errors are comparatively small. Vertical wind profiles could be evaluated against Doppler wind lidar measurements, which were performed in Zurich during a limited period of time between September 2022 and March 2023. (ii) Errors in turbulent mixing, notably due to errors in the stability class selection. A possible way forward could be to use the Eddy flux tower measurements at Hardau to determine atmospheric stability instead of letting the model select the stability class based on the catalogue selection procedure. (iii) Errors due to the discretized nature of the catalogue approach. This was mitigated to some extent by averaging over the five best scoring situations rather than choosing a single one. (iv) Errors due to the resolution not being sufficient to fully resolve the flow in street canyons. The increased deviations of the model from sensors at street level suggest that this is a significant limitation for such sites. (v) Model representation errors. At some sites, especially those at street level, we noticed that sampling the model output only one grid cell (10 m) away or one altitude level higher or lower, significantly altered the results. Capturing the concentrations at sites influenced by nearby sources will remain a challenge. Such sites will likely have to be excluded in inverse modeling. (vi) Errors in background concentrations. Uncertainties in daily afternoon background levels are in the order of 1 to 3 ppm, which is significant compared to the average enhancements from emissions and biospheric fluxes in the city of about 10 ppm. (vii) Sampling noise of the Lagrangian particle dispersion model. Since the concentration in a given volume is determined from the CO_2 mass of all Lagrangian particles in this volume, the statistical noise depends on the number of simulated particles. The selection of a suitable number of particles is a tradeoff between precision and computational cost. Although this noise may be significant at certain locations and times, it does not introduce systematic errors. (viii) Errors due to applying the sector-specific temporal profiles uniformly across the city. These errors depend on the sector and may vary considerably over time. Examples of increased errors are a traffic congestion in one part of a city, or a temporal shutdown of an industrial source. To enhance the flexibility of the model to capture such situations, it would be necessary to perform simulations not only per sector but also per location.

- How should the reader thinking about the various sources of uncertainty in inversions. Presumably this approach largely aims to minimize the model representation error. Can we expect that other sources of error therefore are more important (measurement error, background error)?

Determining realistic uncertainties is crucial for atmospheric inversions. This includes uncertainties in the prior assumptions (emissions, background levels), the measurements, the atmospheric transport model, and the model representation. As this will be addressed in a forthcoming publication, we prefer not to discuss this here. The comparisons between the GRAMM/GRAL forward simulations and the observations presented in Figures 13, 14 and 15 provide an idea of the magnitude of the combined uncertainties. In response to reviewer #2, we added an analysis of the uncertainty associated with background concentrations, which is clearly more important than measurement uncertainty. The inversion will provide further insights into the separate contributions from prior assumptions versus measurement and model errors, but this is outside the scope of the present study.

- Additional discussion of whether a 10 m simulation is necessary for inverse modeling and offers significant benefit beyond lower resolution forward model would be helpful. The diurnal disagreement in Figure 15 is substantial. If this model is not replicating observed concentrations, would a lower resolution model have sufficed for inverse modeling? Some quantification of the improvement in modeled concentration of this model over others would be quite helpful or at least a discussion of how the reader might think about that question.

We have submitted an inverse modeling study using the mesoscale atmospheric transport model ICON-ART at 500 m resolution where we compared the simulations with the same mid-cost CO_2 observations (Ponomarev et al., submitted to Atmos. Chem. Phys.). The range of root-mean-square errors between ICON-ART and the rooftop sensors in Zurich was 10-16 ppm for afternoon mean concentrations. The corresponding values for GRAMM/GRAL are 8-12 ppm, suggesting that the higher resolution is indeed beneficial. However, these differences are not only due to resolution but also due to the largely different modelling approaches. Again, we prefer to discuss this in a forthcoming publication on inverse modeling with GRAMM/GRAL where we will analyze the performance statistics of GRAMM/GRAL in detail and discuss the comparison with ICON-ART.

The diurnal disagreement is indeed substantial, which will likely make it necessary to limit the data assimilation to those hours of the day, where the model performs best. Using only daytime or afternoon observations is common practice in atmospheric inverse modeling because it is much more difficult to reproduce mixing in shallow nocturnal boundary layers than well-mixed daytime atmospheric boundary layers.

- For lines 427 - 438 (and Figure 16): How does the seasonal and diurnal variability in the probability density of stability classes prove that the distribution of stability classes are correct?

We didn’t claim that the distribution is correct, but made a more qualitative statement that there is “no indication for a major misrepresentation of the atmospheric boundary layer dynamics by the model.” The diurnal cycles of the frequency distributions of stability classes in different seasons look plausible as argued in the manuscript, but it is possible that certain classes are over- or underrepresented. One disadvantage of dividing the weather situations into discrete stability classes is that the dispersion changes stepwise and significantly between neighbouring stability classes. As a result, even a small misattribution (shift by 1 stability class) can have significant consequences. We added the following sentence to this point:

Nevertheless, certain stability classes may be under- or overrepresented, which can have significant consequences. As demonstrated by Ars et al. (2017) and Hanfland et al. (2022), a small shift by only one stability class can lead to differences in peak concentrations and volumes of plumes from individual emission sources by up to a factor of two.

- I suggest adding an additional paragraph addressing the implications of this paper at the end.

The conclusions indeed ended too abruptly. We added the following lines at the end:

This study highlights the potential of high-resolution, building-resolving atmospheric modeling frameworks like GRAMM/GRAL to accurately simulate urban CO_2 concentrations when paired with detailed emissions inventories and biospheric flux modeling. Its ability to support inverse modeling and provide near-real-time insights into emission patterns can greatly enhance transparency and accountability for cities pursuing net-zero goals, and it offers a scalable blueprint for other urban areas globally.

Line-by-Line Edits

- Figure text is too small throughout the manuscript

We have increased the size of labels, legends and titles for all figures.

- What is meant by the last line of Table 1? The line with “Example” and “how we should do.” Final line is an “example” and should be removed.

Thank you, that was a mistake. The line should have been uncommented.

- Description Equation 1, for the unfamiliar reader, describing the units of each term would be helpful for understanding the equation

Concentrations are provided in ppm (dry air mole fraction) throughout the manuscript, since this is the units in which all measurements are reported. The temporal scaling factor tau has no units. We added this information.

- Figure 11, is the white in the city center missing data due to the building height higher than 2 m?

Yes, this is the reason. We added this information to the figure caption.

- Line 36: It is unclear what “suitable measurement” means.

We agree. The sentence was reformulated as follows:

Emissions from a city produce a gradient in atmospheric concentrations between regions upstream and downstream, which can be measured by a measurement network with stations suitably placed to capture these gradients.

- Figure 5: Color key shows values of “-999” to “0.” The value -999 must be a filler for unknown values, so it should be removed or explained.

We added the following information to the figure caption:

White color indicates no information for the corresponding building block or zero occupancy.

- Figure 7: Images b and c require x axis labels.

We think that the labels (‘Mon’, ‘Tue’, etc. in Fig.7b and ‘Jan’, ‘Feb’, etc. in Fig.7c) together with the figure caption are sufficiently clear.

- Figure 8: How were the different vegetation patches selected as representative?

This was entirely based on visual inspection looking for areas with very homogeneous coverage by the corresponding vegetation (using information on vegetation type provided by the data sets listed in Table 2 as well as aerial satellite imagery).

- Table 3: How was the threshold determined for match2obs selected sites.

We did not apply a threshold to determine if a site was selected or not. As explained in the text, we excluded sites not located on rooftops and, as in the case of Uetliberg, on a high mountain. We also excluded the rooftop sites Zürich Irchel and Bankenviertel where the wind sensors were placed too close to the roof or building because of logistic constraints. We provide further information on the reason for excluding Uetliberg in a response to a comment of reviewer #2.