Reviewer 1

Overall Comments

This manuscript addresses a topic that is both timely and scientifically important. The use of TROPOMI NO2 observations within an ensemble Kalman filter framework to infer European fossil fuel CO2 emissions is a promising direction, and the reduced-complexity NOx chemistry module is a useful technical contribution. I appreciate the authors' transparency in acknowledging several key limitations.

That said, I have some concerns about the framing and interpretation of the results. Some parts of the manuscript seem to overstate the capability of the proposed technique, while the methodology as presented seems to be at the proof-of-concept demonstration stage. This is not a criticism of the science itself, and I think the manuscript would benefit from more clearly reflecting the current limitations and the gap between what is demonstrated here and what would be needed for practical operational application. In particular, the assumed temporally constant NOx:CO2 emission ratios across sectors, seasons, and countries is a significant simplification whose implications for the posterior CO2 estimates are not fully explored. The authors do acknowledge this limitation toward the end of the paper, but given how central it is to the reliability of the results, it deserves more discussion and supporting analyses.

I am also concerned about the large posterior adjustments in national ffCO2 emissions, which range from 20% to 91% relative to the prior. While I recognize that bottom-up inventories can carry significant uncertainties, adjustments of this magnitude at national scales are surprising for a state-of-the-art inventory used in this study, and the manuscript currently provides no way to assess whether these reflect genuine inventory biases or artifacts of the inversion framework itself. I think this finding deserves more careful and critical discussion, including comparison against independent bottom-up emission estimates or other top-down inversion studies over Europe. Without such benchmarking, it is difficult to evaluate the reliability of the main quantitative results of the paper.

I recommend major revision before this manuscript is suitable for publication.

We thank the reviewer for their careful and constructive assessment of our manuscript, and for highlighting both the timeliness of the topic and the value of the reduced-complexity NOx chemistry framework. We appreciate the recognition of the scientific potential of this approach.

We also acknowledge the reviewer's concerns regarding the framing and interpretation of the results. We agree that the current study does represent a proof-of-concept demonstration rather than a finalised methodology or operational system. In response, we have revised the manuscript to more clearly reflect this positioning. Specifically, we have moderated statements in the abstract, introduction, and conclusions that could be interpreted as overstating the capability of the technique, and we now emphasise the methodological development and its current limitations more explicitly.

We have expanded the discussion of the assumptions underlying the NOx-to-CO2 inference, particularly the use of temporally fixed prescribed NOx:CO2 emission ratios. We clarify that CO2 emissions are inferred diagnostically from posterior NOx fluxes using these fixed ratios, rather than being directly optimised within the EnKF framework. We further discuss how variability in these ratios across sectors, regions, and seasons may introduce systematic biases in the inferred CO2 emissions, and we highlight this as a central limitation of the present approach.

We also agree that the large posterior adjustments in national ff CO2 emissions (20–91%) are a key result that requires more careful interpretation. In the revised manuscript, we have expanded the discussion of these changes to more critically examine their possible origins. While such adjustments could partly reflect biases in the prior inventory, we emphasise that they may also arise from structural limitations of the inversion framework, most significantly related to assumptions in emission ratios, as well as limitations of the inversion framework (single day assimilation, fixed non-combustion fluxes at prior, fixed boundary conditions), and potential impact of simplifications in the offline NOx chemistry.

To better contextualise these results, we have included a comparison with EDGAR emissions, highlighting that the magnitude of the inferred national changes sometimes largely exceeds typical inter-inventory differences. We therefore explicitly state that these results should be interpreted with caution and that the current framework does not yet provide a fully robust constraint on national CO2 emissions.

Overall, these revisions aim to provide a more balanced and transparent interpretation of the results, clearly distinguishing between what is demonstrated in this study and what remains to be addressed in future developments toward operational applications.

Specific Comments

Lines 6-7: “We estimate European NOx emissions for 2021 by assimilating TROPOMI NO2 observations into an Ensemble Kalman Filter (EnKF) framework, optimised within the GEOS-Chem atmospheric transport model.”

Although this statement is true, the manuscript never presented posterior NOx emissions. Consider adding a new figure dedicated for NOx emissions or add NOx layer onto an existing figure (i.e., Figure 6).

The NOx emission adjustments have a very similar pattern to the presented CO2 changes, due to the fixed emission ratio used to derive ffCO2. However, we appreciate this would be helpful to show as an integral part of the inversion process. We have now included a temporal plot of the total European NOx prior and posterior fluxes as part of Figure 6a. Additionally, the individual country NOx posterior and prior are shown in appendix, Figure A2.

Lines 20–22: "Accurately monitoring these emissions is essential for tracking progress toward national mitigation goals, yet it is complicated by the difficulty of isolating fossil fuel signals from large, seasonally varying biogenic fluxes."

As written, this sentence implies that atmospheric observation-based monitoring is the primary or only method for tracking mitigation progress. In practice, self-reported national inventories, global bottom-up emission datasets, and other approaches all contribute to this goal. I suggest revising to acknowledge the broader monitoring landscape and to clarify that the atmospheric top-down approach complements rather than replaces these efforts.

We thank the reviewer for this comment and agree that atmospheric methods are one component of a broader emissions monitoring framework. We have revised the sentence in the Introduction to clarify that national inventories and bottom-up datasets provide the primary basis for tracking mitigation progress, while atmospheric top-down approaches provide an independent and complementary constraint.

Figure 1b: The blue (ICOS) and green (DECC) triangles are difficult to distinguish visually. Please consider changing the color or marker style of one of the two site types.

We have adjusted this to better contrasting colours and used different marker styles for site type (as response to below comment).

Figure 1b (continued): One of the NOAA in situ CO2 measurement sites appear to be located over a marine area. Could the authors confirm whether this is the Ocean Station Norway (STM) site? If so, my understanding is that observations at this site were terminated in 2009, and it would be important to clarify which dataset is actually being used here.

We thank the reviewer for identifying this issue. Upon review, we confirmed that the site labelled as NOAA corresponds to the Shetland Islands (SIS) site, which is part of the ICOS network. Similarly, the other site labelled as NOAA is Observatoire de Haute-Provence (OHP), which also contributes to ICOS. We apologise for this mislabelling and have now corrected the figure and manuscript text to reflect these classifications accurately.

Lines 84–91: The authors describe the EEA NO2 network and the ICOS/DECC/NOAA CO2 sites used for evaluation, but there is no discussion of site representativeness. Many ICOS and NOAA tall tower sites are specifically designed to sample regional background air, which may make them poorly suited for detecting near-source ffCO2 emission changes. Similarly, some EEA NO2 sites may be located far from major

combustion sources. Please add a brief characterization of the site types used here — urban, suburban, rural background — and discuss how their representativeness affects the interpretation of the model evaluation results.

This is an important point, and worth expanding upon in our discussion. We have now categorised all the situ CO2 sites into urban, suburban, and rural (shown in Figure 1 by marker style). The majority of the CO2 monitoring sites are rural background stations (all DECC, and 38 of the 42 ICOS sites), reflecting their design to sample regional background air. In contrast, the EEA NO2 sites are more focused on urban environments, as they are primarily intended to monitor population exposure and emissions from traffic and combustion sources. Given the large number of EEA sites (>200), it was not feasible to manually categorise each one, and no consistent classification is provided within the EEA dataset. However, we expect the network to be dominated by urban and suburban sites, with rural background sites forming a minority.

We have expanded the discussion in Section 3.2 to clarify how site representativeness affects the interpretation of the model evaluation. In particular, we note that urban near-source NO2 measurements are sensitive to sub-grid-scale emission plumes, contributing to the negative bias in model–observation comparisons, while the predominantly rural CO2 network is less sensitive to local fossil fuel emission changes, which can help explain the more limited improvement in CO2-based evaluation metrics.

Lines 101–103: "Prior combustion emissions of NOx and CO2 are taken from the CAMS-REG v8.1 emissions inventory at 0.05°×0.1° resolution."

Given how central this dataset is to the entire study, a brief description of how these emissions are estimated would be helpful — for example, whether they are based on national inventory reports, fuel consumption statistics, or activity data. Also, the CO2 and NOx maps in Figure 2 appear to show patterns that align with national boundaries. Is this a result of aggregating independent national inventories? If so, what are the implications for the inversion results? (i.e., do uncertainties vary by country?)

We have added additional detail describing how the CAMS-REG inventory is constructed, including its basis in national emission inventories, activity statistics, and emission factors, together with spatial disaggregation using proxy datasets. The reviewer is right that the spatial patterns in the emissions reflect the combination of independently reported national inventories, which can lead to visible discontinuities across country borders. We have clarified this in the manuscript and note that such structure is an inherent feature of bottom-up inventories.

We also expanded the discussion of the associated uncertainty framework. The CAMS-REG uncertainty product (Super et al., 2026) provides spatially resolved, sector-dependent uncertainties that vary by country, reflecting differences in reporting methodologies and data quality. These spatial variations in prior uncertainty directly influence the inversion, as they determine the allowable magnitude of posterior adjustments in different regions. We have added a short discussion of these implications in the Methods section.

Lines 105–106: "Emissions from non-combustion sources in the CAMS-REG v8.1 inventory — which include fugitives, waste, solvents, and agriculture — are kept fixed at their prior values."

What is the justification for fixing non-combustion NOx emission sources while only adjusting combustion sources? This decision deserves more discussion. In particular, biogenic soil NOx emissions can represent a significant fraction of total NOx over agricultural area during summer, and fixing these at potentially biased values could systematically drive spurious adjustments in the posterior combustion NOx estimates. Please justify this choice more explicitly or acknowledge it as a limitation with a discussion of its potential seasonal implications.

In this study, we restrict the inversion adjustments to combustion-related NOX emissions, as these sources are directly linked to fossil fuel CO2 emissions and are therefore consistent with the cross-species constraint employed in the inversion framework. In contrast, non-combustion NOX sources (e.g. soil, waste, agriculture etc.) are not coupled to CO2 emissions and are expected to exhibit different temporal and spatial variability, making them less suitable for joint optimisation in this framework.

We acknowledge that this assumption introduces a limitation, especially for biogenic NOx emissions, which can have seasonal variation over agricultural regions, as the reviewer mentions. In such cases, biases in the prescribed non-

combustion emissions could be partially compensated by adjustments to combustion NOx emissions in the inversion. We have added a discussion in the manuscript to clarify this limitation and highlight its potential seasonal implications.

Lines 110–114: The offline NOx chemistry parameterization is described and validated in Schooling et al. (2025) for emission perturbations on the order of +-20%, If I understand correctly. However, the posterior results reported in Section 3.1 show adjustments of up to +91% at the national scale and widespread winter increases of 10–177% at the grid level. It is unclear whether perturbations of this magnitude fall within the validated range of the linear scaling approximation. Please add a discussion of the expected error in the offline chemistry scheme under these larger perturbations, and assess whether this could introduce systematic biases in the posterior results.

Lines 123–125: The stability of the NO2:NOx partitioning ratio under emission perturbations is cited as a key assumption enabling the conversion of modeled NOx to NO2 columns for comparison with TROPOMI. However, similar to the point above, the range of perturbation magnitudes over which this stability holds is not discussed. What is the uncertainty introduced by this assumption, and how is it accounted for in the inversion framework? Given the large posterior adjustments found in this study, this assumption warrants more explicit validation or at minimum a sensitivity discussion.

We thank the reviewer for this important comment. While the offline NO_x chemistry scheme was originally validated for perturbations of order ±20% (Schooling et al., 2025), we recognise that the posterior adjustments in this study are larger. To address this, we have performed additional tests using the GEOS-Chem full chemistry to assess the validity of the linear scaling approximation over the range of perturbations encountered here.

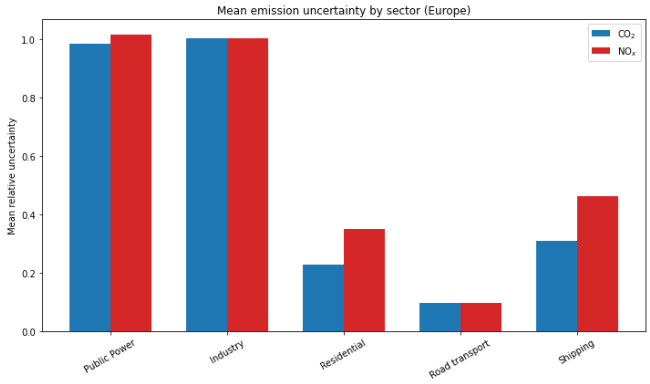
Specifically, we evaluate the offline chemistry scheme for representative days in each season (January, April, July, and October), using emission perturbations consistent with those found in the posterior solution. The results of this analysis are presented in Appendix A1. We find that the offline parameterisation reproduces the full chemistry NOx loss rates with high fidelity, with coefficients of determination (R^2) exceeding 0.94 across all seasons. In addition, the NO2:NOx partitioning remains effectively unchanged under these perturbations (R^2 = 1), indicating that the assumption of stable partitioning remains valid within the range explored.

These results demonstrate that, despite the larger emission perturbations, the linear scaling approximation provides a robust representation of NO_x chemistry for the purposes of this inversion. Any residual errors introduced by the offline approach are therefore expected to be small and are unlikely to introduce systematic biases in the posterior emission estimates. In particular, Schooling et al. (2025) demonstrated that even under conditions where the agreement between offline and full-chemistry representations is weaker than observed here, the resulting impact on simulated NO2 column concentrations is more than an order of magnitude smaller than typical TROPOMI NO2 observational errors. In the context of the present study, where the agreement between offline and full chemistry is stronger ($R^2 > 0.94$), the contribution of the offline chemistry approximation to the total error budget is expected to be minimal.

Line 165: "the values of σn, σc and r are prescribed from the CAMS-REG v8.1 uncertainty estimates and the CORSO cross-species correlation product, and are assumed to be constant in time."

A few questions on this. First, which emission sector carries the largest uncertainty for NOx and CO2 in CAMS-REG v8.1? Based on Figure 3, northern Africa appears to have substantially higher uncertainty for both species, which seems to drive larger error reductions (Figure 4a) and larger CO2 flux increments (Figure 5b) in that region — yet no in situ evaluation is available there. Please discuss the implications of this large uncertainty for the NOx inversion and the derived CO2 estimates over both North Africa and the broader European domain. Second, do NOx and CO2 prior flux uncertainties have seasonality? If so, how would they affect the inversion result? What are the implications for interpreting the seasonality in monthly posterior increment (Figure 5)? Please address this.

The emission sectors with the most significant uncertainty are public power and industry for both CO2 and NOx (see graph).



The reviewer is right to notice the significance in the CAMS-REG uncertainty framework assigning substantially larger uncertainties to regions such as North Africa, reflecting the limited availability of detailed national emission inventories and observational constraints in these areas. As a result, the inversion permits larger emission adjustments there, which explains the more pronounced error reductions and larger CO2 flux increments observed. We acknowledge that in the absence of independent in situ validation data, the robustness of these posterior adjustments over North Africa is more difficult to assess. In such regions, the inversion is more strongly influenced by the prescribed prior uncertainty rather than observational constraints, meaning that large posterior changes should be interpreted with caution.

Over the broader European domain, where prior uncertainties are generally smaller and observational coverage is denser, the inversion is more tightly constrained and posterior adjustments are therefore more robust. We have added a discussion in the manuscript to clarify that regional differences in prior uncertainty can influence both the magnitude and reliability of inferred emission changes, particularly in data-sparse regions such as North Africa.

The CAMS-REG v8.1 uncertainty estimates are provided as annual values and do not explicitly include seasonal variability. In this study, we therefore assume that prior uncertainties for both NOx and CO2 are constant in time. In reality, uncertainties clearly could exhibit seasonal dependence. The lack of seasonal variability in the prescribed uncertainties means that the inversion does not account for potentially larger uncertainties during specific periods (e.g. summer biogenic influences). The inferred seasonal structure in the monthly posterior increments (Figure 6) reflects variations in the model observation agreement (as well as variations in TROPOMI coverage) rather than any changes in the confidence of the prior emissions. A temporally varying uncertainty product would provide a useful additional constraint in future inversion studies.

Figure 3: Please clarify the units of the color bar. Do all four panels share the same units?

The shared colour bar represents relative uncertainty (dimensionless), expressed as the standard deviation normalised by the prior emission magnitude (σ / E). This applies to the NOx and CO2 uncertainty fields as well as the combined prior uncertainty derived from the covariance matrix Pf.

The cross-species error correlation is also dimensionless and lies within [-1, 1], so it is consistent with the shared colour scale. We have clarified this in the figure caption.

Lines 211–212: The authors note that MAE and bias improvements are largest in winter and attribute this to "challenges in bottom-up inventories for domestic heating." While plausible, this explanation is presented as a general conclusion without supporting evidence or citation. Does the prior error covariance have a seasonal component that could contribute to this pattern? Also, the seasonal pattern in prior model vs. TROPOMI agreement (higher errors in winter, lower in summer) is interesting, but is not explained. Please expand this discussion and provide supporting references for the domestic heating hypothesis, if available.

The proposed explanation is supported by previous studies showing that residential combustion emissions are highly seasonal and strongly dependent on meteorological conditions, fuel use, and assumed temporal profiles, making them more uncertain during winter months. These factors can lead to larger discrepancies between bottom-up inventories and observations during periods of peak heating demand. We have added references and expanded the discussion in the manuscript to support this interpretation.

In response to the question on the prior error covariance – this does not include an explicit seasonal component, as uncertainties are prescribed as temporally fixed. However, the observation error covariance matrix **R** is related to the TROPOMI precision values, which does have seasonal variation (often higher in winter, when retrieval conditions worsen with cloud etc). Therefore, the seasonal pattern in posterior adjustments, reflects a combination of (a) seasonal variations in model–observation mismatch and (b) changes in **R**. We find the differences between the prior model and TROPOMI observations are significantly larger in winter, and this is further amplified by longer NOx lifetimes and more stable boundary layer conditions, which increase the sensitivity of NO2 columns to emission errors.

Lines 214–215: It is unclear how the posterior ffCO2 fluxes are actually computed from the posterior NOx scaling factors. The current methods section does not include an explicit description of this step. I strongly recommend adding a dedicated subsection describing the NOx-to-CO2 conversion. Specifically: is the same scalar scaling factor applied uniformly to all sectors at each grid cell? What are the implications of this approach when different sectors co-locate on the same grid cell with very different NOx:CO2 ratios? This NOx-to-CO2 conversion works when the inventory error originates from activity data but breaks down when the error comes from emission factors. This distinction is important for interpreting the posterior CO2 results and should be made explicit in the paper.

We thank the reviewer for this important point. We have added a dedicated subsection to describe of the NOx-to-CO_2 conversion in the Methods section. In our framework, posterior CO2 fluxes are derived by applying the optimised NOx scaling factors to co-located CO2 emissions, assuming a fixed NOx:CO2 emission ratio at each grid cell. This ratio is prescribed from the prior inventory and reflects the relative contributions of different combustion sectors within each grid cell.

This approach is consistent with the prior error covariance, which represents aggregated multi-sector uncertainties at the grid scale. However, a single scaling factor is applied per grid cell, so sector-specific errors are not explicitly resolved. As such, the method assumes that NOx and CO2 errors are primarily driven by shared activity uncertainties. We have clarified these assumptions in the manuscript, making it clear that our methodology using prescribed emission ratio to convert to ffCO2 is a key limitation of this study.

Lines 223–226: Please clarify whether "emissions" here refers specifically to ffCO2 emissions or NOx emissions, as the text could be read either way.

This is ffCO2, we have now explicitly stated this.

Figure 6: A few comments. First, the log-scale y-axis for CO2 flux makes it difficult to visually assess the magnitude of prior-to-posterior adjustments across countries. For example, the corrections for Turkey and Sweden appear visually similar to those for other countries despite being much larger in relative terms. Unless the log-scale axis is required for some reason, consider using either a linear scale or showing the relative adjustment explicitly. Second, for readers unfamiliar with European geography, adding country labels to at least one of the map figures (Figures 4 and 5) would make it much easier to connect the spatial patterns in those maps to the country-level results in Figure 6.

We have changed to non log-scale, and added country labels to the first map in figure 4.

On the large posterior adjustments (Lines 229–231): The reported annual increases of 20–91% across all analyzed countries are surprising and require deeper discussion. CAMS-REG v8.1 is a state-of-the-art inventory, and systematic errors of this magnitude would be unusual and would likely have been identified in prior literature. Changes exceeding 6σ of prior uncertainty for Spain and Turkey should prompt careful examination of whether the inversion framework is functioning as intended — including possible issues with boundary condition, NOx chemistry, NOx:CO2 ratio assumptions under large perturbations. Please provide any independent evidence or plausible physical hypothesis that could support adjustments of this magnitude, and discuss alternative explanations more explicitly. A comparison against other bottom-up inventories (i.e., EDGAR) would provide additional context to this result.

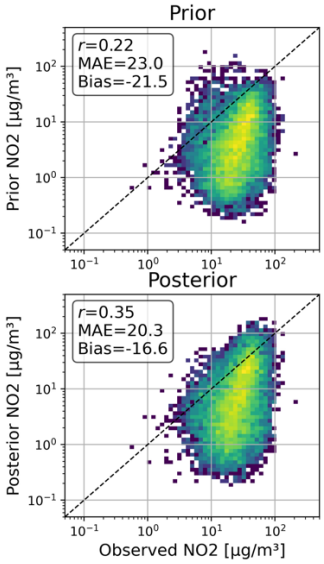
We agree that some of the posterior adjustments are surprisingly large, and acknowledge that there are a number of underlying assumptions and limitations to our inversion framework. We have added a more explicit and detailed discussion of these limitations in the manuscript, stating that the inferred changes here should not be interpreted as definitive corrections to national emissions.

We have also included a comparison to EDGAR for the annual ffCO2 emissions for each of the 12 top emitting countries (Figure 8). Interestingly, we do see improved agreement with EDGAR due to our estimated increases relative to the prior in 5 of the 12 countries (Sweden, Romania, Turkey, Spain, and UK) – most notably the huge posterior change in Turkey reflects much better agreement with EDGAR. However, the remaining 7 countries demonstrate an increase relative to the prior that worsens agreement with EDGAR. In particular, the increases in Germany, France, and Italy show considerable (>100,000 Mt) increase relative to both of the independent bottom-up datasets and should therefore be interpreted with caution.

On the large posterior adjustments and surface bias (Figure 7): Relatedly, the posterior barely improves the surface NO2 bias (-18.8 to -16.4 ug/m^3) despite national emission increases of 20–91%. This is quite surprising and is not directly addressed. How would the evaluation look if restricted to in situ sites located in or near the countries with the largest adjustments (i.e., near cities or combustion sources)? This analysis would help assess how the posterior emission increases are reflected in near-surface concentrations.

Restricting the analysis to 35 of the EEA sites, which are in close proximity to regions of high NOx combustion flux, and therefore more sensitive to the emissions adjustment we find the following. In these sites the negative bias is even worse in the prior (-21.5), but very similar in the posterior (-16.6) – reflecting a more substantial improvement in these regions. However, the general pattern of agreement is very similar in this small subset compared to the whole network, so we don't deem this plot worthwhile including in the manuscript as there isn't much new information to gain.

We believe the reason for this negative bias is largely related to the coarse model grid size (~25 km) in GEOSChem which will dilute the magnitude of concentrations in close proximity to point source emissions. We have included more emphasis on this in the discussion of our results (as also addressed in your previous comment relating to the region type of the different in-situ sites).



On the temperature–ffCO2 relationship: The negative correlation between ffCO2 flux and surface temperature is a physically intuitive and visually compelling result. This relationship is one of the key scientific findings of the paper and would benefit from additional support, such as from bottom-up inventory analyses. A potential point of discussion: the temporal emission profiles in the prior (Figure 2, bottom row) already encode higher winter emissions for heating sectors. If the inversion is primarily amplifying a seasonal signal that was already present in the prior, the strengthened temperature-flux relationship in the posterior may partly reflect prior structure being scaled up rather than genuinely new information from TROPOMI? Also, how should energy demand for summer cooling (air conditioning) affect the ffCO2–temperature relationship?

The prior temporal profiles already impose a seasonal structure in ffCO2 emissions, particularly for heating-related sectors, and this contributes to the baseline temperature–ffCO2 relationship. However, the inversion does not introduce this relationship a priori, but rather adjusts its magnitude in response to the observed NO2 constraints. The strengthened temperature–flux relationship in the posterior therefore reflects both the prior structure and additional

information from TROPOMI, which acts to amplify, dampen, or redistribute emissions where model–observation mismatches are largest. In particular, the largest posterior adjustments occur during winter months, when both emissions and NOx sensitivities are highest, indicating that the inversion is actively correcting biases rather than simply scaling a fixed seasonal pattern. At the same time, uncertainties in the prescribed NOx:CO2 emission ratios for these sectors, particularly for residential combustion, may lead to an overestimation of the inferred ffCO2 response if NOx variability is not matched by proportional CO22 variability.

Regarding summer conditions, we note that increased energy demand associated with cooling can contribute to ffCO2 emissions at higher temperatures, particularly in the power sector. However, this signal is generally weaker and more spatially heterogeneous than winter heating demand in the European domain. Electricity-driven summer cooling emission sources may contribute to CO2 without such a strong NO_x signature. This is especially relevant where cooling demand is met by relatively low-NOx or cleaner electricity generation, reducing the apparent coupling between temperature and emissions. We have expanded the discussion of our temperature results to include these nuances.

On figure A2 and the weekday-weekend signal: The appendix figure showing that the prior has two distinct branches in the temperature-flux scatter — corresponding to weekday and weekend emission regimes — that merge into a single coherent relationship in the posterior is an interesting finding that deserves more prominence and discussion in the main text. Is this merging physically meaningful, suggesting the inversion is correctly capturing a smoothed emission signal? Or could it reflect the Savitzky-Golay smoothing washing out the existing weekly cycle? It would also be informative to show the equivalent NOx flux vs. temperature scatter plot (a NOx version of Figure 6) to assess whether the same pattern holds for the directly constrained species.

The two distinct branches in the prior reflect the fact that emissions are prescribed with a fixed weekly cycle and no week-to-week variability within a given month, leading to distinct weekday and weekend regimes in the temperature–flux relationship. In the posterior, the weekly cycle (with drops each weekend) remains present (as seen in Figure 6a) and is not removed by the Savitzky–Golay smoothing. However, the inversion introduces substantial day-to-day and week-to-week variability in emissions in response to observational constraints. This added variability leads to a more continuous and coherent temperature–flux relationship, rather than two discrete branches. We interpret this as a physically meaningful result, indicating that the inversion relaxes the rigid temporal structure of the prior and produces emissions that are more consistent with real-world variability in activity and energy demand.

We have also included the NOx version of Figure 7 in the appendix, which demonstrates a similar pattern with temperature.

On the OCO-2 evaluation (Figure 9): The near-absence of improvement in the OCO-2 comparison is acknowledged and expected given the dominance of biogenic signals in XCO2. However, the positive OCO-2 model bias that worsens in autumn and winter in the posterior is worth examining more carefully — what does this imply about the biospheric fluxes used in the inversion? More broadly, including OCO-2 as an evaluation dataset when it demonstrably cannot detect the signal being estimated risks creating a misleading impression of multi-dataset validation. Please either clarify the specific role and limitations of the OCO-2 evaluation more explicitly, or reconsider whether it adds sufficient scientific value to warrant inclusion.

We agree that OCO-2 provides limited sensitivity to fossil fuel CO2 signals, and for this reason cannot be used as a meaningful independent validation of the inversion. Rather, the purpose of this comparison is to assess whether large posterior adjustments to ffCO2 emissions are detectable in column CO2 observations. The lack of a clear improvement or any significant change in the OCO-2 comparison, despite substantial changes in the inferred combustion fluxes, highlights the limited sensitivity of XCO2 to fossil fuel emissions in regions where biospheric fluxes dominate. We consider this to be an important result in itself, supporting the broader motivation for using short-lived tracers such as NO2, which provide greater sensitivity to combustion sources in top-down inversions.

The persistent positive model bias, which is worsened in the posterior state in autumn and winter, is likely to be driven primarily by uncertainties in biospheric fluxes, as well as boundary and initial conditions to which CO2 is particularly sensitive. Seasonal biases in ecosystem respiration and uptake during these periods can dominate the total column signal, offsetting the relatively small fossil fuel contribution. However, it is clearly difficult to isolate the contribution of combustion fluxes to the model error. Future work combining NO2 and CO2 observations within a single inversion framework could help to disentangle biogenic and anthropogenic contributions and better constrain both flux components simultaneously. We have revised the manuscript to clarify the role and limitations of the OCO-2 evaluation and to expand the discussion of these points.

On spatial evaluation: The model evaluation is conducted as temporally aggregated statistics across all available sites. Given that the inversion produces different results across countries — large increases in Turkey and Spain, modest changes in Netherlands and Belgium — a spatially resolved evaluation would provide much more diagnostic insight than domain-wide statistics alone. Or even a simple map of site-level prior vs. posterior would help assess whether the regional patterns in the posterior are supported by independent observations.

We have now included an additional figure presenting this analysis (Figure 11). We show the change in model-observation MAE (posterior – prior) for each site for both the in-situ NO2 and CO2 network. We demonstrate the widespread improvement in the NO2 agreement with only a handful of sites in the summer showing a worsening. Conversely, we show how the changes with CO2 are generally more variable across the region.

On the TROPOMI overpass time: TROPOMI observes at approximately 13:30 local time, meaning the inversion effectively constrains a single daily snapshot of emissions. The extrapolation to full daily totals relies on the sector-based diurnal scaling factors, and any errors in the assumed diurnal profile shape propagate directly into the posterior flux estimates. Please add a discussion of this limitation and its implications, and consider how complementary approaches such as geostationary satellite observations (e.g., TEMPO, which is directly relevant here), ground-based monitoring networks, or bottom-up activity data could help constrain the diurnal cycle more robustly.

This is a good point, worth acknowledging. It is true that the single TROPOMI overpass means the inversion constrains emissions at this snapshot in time and is not able to infer the diurnal cycle. The extrapolation to daily total emissions relies on prescribed sector-specific diurnal profiles, and uncertainties in these profiles will propagate directly into the inferred fluxes. We have added a discussion in the manuscript to clarify this limitation and its implications.

On the broader methodological context: Several complementary approaches have emerged in recent years for inferring ffCO2 emissions from satellite observations, including direct plume detection methods, machine learning-based approaches, and observed cross-tracer ratio techniques. A brief discussion situating the present approach within this landscape — including its relative strengths and limitations compared to these alternatives — would help readers better understand the contribution and applicability of the method.

We have added to the concluding remarks discussing complementary approaches in the context of our study.

Reviewer 2

The manuscript "Inferring European Fossil Fuel CO2 Emissions using TROPOMI NO2 Data and Sector-Based NOx:CO2 Emission Ratios" by Chlöe N. Schooling et al. documents European scale atmospheric inversions assimilating TROPOMI NO2 observations to derive CO2 fossil fuel emissions maps and national budgets for the year 2021. In the Bayesian framework of the inversion, the assimilation of the NO2 data to infer NOx emissions relies on a simplification of the chemistry modeling in the GEOS-Chem model, and the transfer of information from the TROPOMI NO2 observations to the CO2 fossil fuel emissions is based on a joint correction of the prior NOx and CO2 anthropogenic emission estimates from the CAMS-REG inventory with cross species correlations in the covariance of the uncertainties in these prior estimates.

The topic is important for the atmospheric inversion community. Overall, the inversion and experimental approaches are relevant and interesting, and the text is often clean and concise.

We thank the reviewer for this positive summary and for recognising the relevance and clarity of our approach to this important problem.

However, the text can also go too fast and it can easily be confusing:

- the way the information is transferred from NOx to CO2 emissions should be better explained with a better presentation of its principle in the introduction and a clearer section 2.3. The title, the abstract (line 11), the introduction (l33 and l54-55), section 2.3 (up to line 137) and the conclusion (l279-280) implicitly let the reader think that it is based on fixed sectoral NOx to FFCO2 emission ratios. The details of section 2.3 from line 152 show that it actually relies on the correlations of the uncertainties in the prior emission estimates between the two species, which is different. This topic deserves a better presentation of its principle in the introduction. More generally, the principles of the "coupled" inversion framework should be better introduced in the introduction or early in section 2 before entering into the specific parameters of the GEOS-Chem and EnKF configurations.

We can appreciate that the description of the NOx–CO2 relationship required clarification, particularly to avoid the impression that the framework constitutes a fully coupled inversion. We have added a new section (2.4) to the methods to give space to describe this in more detail.

We emphasise that the inversion implemented in this study is not a coupled NO2–CO2 inversion. The EnKF framework directly assimilates NO2 observations and optimises NOx emissions only. CO2 emissions are not part of the inversion state vector and are not independently constrained by observations.

Instead, ffCO2 emissions are inferred diagnostically from the posterior NOx fluxes using prescribed sector- and region-dependent NOx:CO2 emission ratios. In this sense, the mean state relationship between NOx and CO2 is defined by these emission ratios. However, the prior error covariance matrix used in the inversion includes cross-species correlations between NOx and CO2 emissions, reflecting their common combustion sources. As a result, the uncertainty in the inferred CO2 emissions is consistent with the prior uncertainty structure assumed in the inversion. In particular, uncertainties in NOx emissions, as well as their statistical correlation with CO2 emissions, are propagated through to the inferred ffCO2 fluxes.

We have clarified this distinction throughout the manuscript to emphasise how the relationship between NOx and CO2 is constrained.

- the presentation of the diagnostics and figures should be more rigorous: e.g. the statistics at lines 9-11 do not correspond to a well defined set of variables (to well defined spatial and temporal scales for the values and averaging), and there are residual questions regarding this when reaching line 203 (do we have error reduction for emissions averaged over Europe -"Europe wide"- and seasons, or the average over seasons and Europe -as implied by Figure 4a- of the error reduction at the model spatial resolution and daily scale ?). The same type of comment apply to the correlation increase at line 11.

We can see that the description of these statistics in the abstract is not clear. We have described the method for calculating seasonal error reduction more clearly updated the stats used in the abstract.

The legend of Figure 2 does not specify the spatial resolution of the averaged maps. The legend of figure 5a is "full year total European fluxes" but it shows daily emissions. The legend of figure 6 is incomplete... This is just a small sample of examples illustrating a more general problem: there is a need to carefully check all the legends of figures and all the descriptions of statistics in the text.

We have added detail to all captions as necessary.

Regarding the results, analysis and discussions, my main concerns relate to the fit to the TROPOMI NO2 observations and to the evaluation of the FFCO2 emissions resulting from the inversions:

- Line 275 acknowledges that the "posterior flux adjustments imply substantial increases in national ffCO2 emissions relative to the prior inventory" but section 3 does not highlight it enough. Emissions of the most emitting countries in continental Europe, i.e. Germany, Italy and France, are increased by 36 to 48% at annual scale, and the emissions of the other countries within the largest regional emission cluster (underlying the highest signal in the TROPOMI NO2 images), i.e. Netherlands and Belgium, are increased by ~30%. From that point of view, it is difficult to follow the general statement of "improved ffCO2 estimates" (line 12), lines 15-17 ("successfully isolates the fossil fuel component") and more generally the abstract, which ignores these very large increases in national budgets, and the conclusion at lines 273-274.

We thank the reviewer for this important comment. We agree that the magnitude of the inferred posterior emission increases for major emitting warrants clearer emphasis and contextualisation in the manuscript.

In response, we have revised Section 3, the abstract, and the conclusions to more explicitly highlight the magnitude of these changes and to ensure they are not overlooked by the reader. We now clearly state that posterior adjustments imply substantial increases in national emissions, in some cases exceeding prior uncertainties and the variability between independent bottom-up estimates. We emphasise that these results should be interpreted with caution.

To provide additional context, we include a comparison with independent bottom-up estimates from EDGAR. While the posterior increases are large, we find that for five countries they lead to improved agreement with EDGAR, suggesting that part of the adjustment may reflect biases in the prior inventory. However, we also acknowledge that for other countries the agreement deteriorates, highlighting potential structural limitations in the inversion framework.

We have strengthened the discussion of these limitations throughout the manuscript, including uncertainties in emission ratios, fixed non-combustion NOx sources, assumed diurnal emission profiles, and the absence of direct CO2 constraints. We now frame the study more explicitly as a proof-of-concept methodology, and emphasise that the inferred emission changes should not be interpreted as definitive corrections to national inventories, but rather as an illustration of the sensitivity of tracers such as NO_2 to combustion-related emissions.

Finally, we have revised the abstract and key statements to avoid over-generalised language such as “improved ffCO2 estimates” without qualification, and instead explicitly state the large increases found in national budgets and emphasise the conditional and exploratory nature of the results.

- Surprisingly, these large increases correspond to small decreases of the statistical uncertainties in the NOx emission estimates (figure 4a) and to marginal improvements of the fit to the TROPOMI NO2 data (figure 4b; lines 206-213 should better acknowledge the weakness of these improvements).

The relatively modest reduction in posterior uncertainty (mean of ~4–6% at grid-cell level) despite substantial emission adjustments reflects several factors inherent to the inversion framework. First, uncertainty reduction depends on the ensemble spread rather than the magnitude of the posterior increment itself; large emission adjustments can occur as coherent shifts of the ensemble mean without significantly reducing ensemble variance. Second, the effective reduction is constrained by the limited observational information content, such that only the component of the prior uncertainty that projects onto observation space can be reduced. Third, the use of spatially correlated prior errors reduces the number of independent degrees of freedom, limiting the achievable variance reduction even when local adjustments are large. In addition, diagnostic aggregation of uncertainties assumes spatial independence, and therefore yields a lower-bound estimate of uncertainty reduction at large scales. These combined effects explain why grid-scale reductions appear modest, even though local adjustments can be large and the total European flux uncertainty shows a more substantial reduction (of ~40% from 5.6% to 3.3%).

We agree that the relatively modest improvements in model–observation agreement, despite substantial posterior emission adjustments, also require further clarification.

First, we note that while the improvements in aggregated metrics appear modest, the inversion produces a consistent and systematic response across the full dataset. This is illustrated more clearly in the revised manuscript with the addition of Figure 5b, which shows a near-linear relationship between prior model error (observation–prior) and model change (posterior–prior) at the TROPOMI observation level. This shows the relationship between prior error (obs-pri) and change in model (post-pri) in NO2 column observation space. Here, any data that falls on the line, y=x, represents a point that has been perfectly adjusted to agree with the observation. From this plot, it is clear that the majority of the data is clustered around this line, and apart from a few outliers we have a trend of posterior change having the same sign as prior error, meaning we are correctly compensating to give improved agreement.

There are some data where prior error has a large positive value (where the model is under-estimating observations) but the posterior change to the columns is much smaller in magnitude, and to a smaller extent data with significant negative value, but much smaller posterior changes. These represent the areas where the columns are relatively insensitive to changes in combustion fluxes. In particular, the model–observation mismatch is influenced not only by emission uncertainties, but also by background concentrations, which depend strongly on boundary conditions and initial (restart) fields. Additionally, any uncertainty in the fixed biogenic and non-combustion NO2 sources mean some areas may have model errors which are uncompensated for due to insensitivity to the combustion flux adjustments. These components are not optimised in the inversion and therefore constrain the extent to which emission adjustments alone can improve agreement with observations, especially in regions dominated by background concentrations or weak sensitivity to surface emissions.

In addition, the inversion is performed using a one-day assimilation window, which restricts the temporal propagation of information. Given the short lifetime of NOx and the reliance on a single daily satellite overpass, observational sensitivity is limited to emissions within a narrow time window. This reduces the ability of the inversion to accumulate corrections over time and damp persistent model biases. Future work could address these limitations by extending the inversion framework to include temporal error correlations and multi-day assimilation windows, allowing information to propagate more effectively in time. Additionally, explicitly accounting for uncertainties in boundary conditions or jointly optimising background concentrations (for example with additional ground-based measurements), as well as incorporating biogenic fluxes in the state vector could improve the overall fit to observations.

We have revised the manuscript to discuss the reasons for limited improvement in observation agreement in some areas.

This raises major questions, and this probably requires more extensive discussions and new diagnostics:

- How can we explain that the improvement of the fit to TROPOMI NO2 observations is very limited despite very large corrections to the prior estimate of the emissions (at least of the FFCO2 emissions) far beyond their prior uncertainties at daily scale ? Does the EnKF framework allow for comparing the projection of prior and posterior uncertainties in the observation field to the observation errors and could it help answer this question ?

We would disagree that the improvement in fit with TROPOMI being very limited. We hope that our addition of Figure 5b, which presents this more clearly for the full year of data, that we can convince the reviewer that for the majority of assimilated data we see good improvement in posterior relative to prior fit. However, we do acknowledge that the posterior agreement is not perfect and can be very limited in regions that are less sensitive to combustion flux. As mentioned in the previous response, we have now added a discussion of the limitations on potential areas this could be expanded on in future work.

There is not a one-to-one correspondence between large emission adjustments and improvements in the fit to TROPOMI NO_2 observations due to fundamental limitations in what the EnKF framework can constrain. This reflects the fact that only the component of the prior error that projects onto observation space through the model operator can be corrected. Errors that lie outside this subspace (e.g. transport errors or background biases) cannot be reduced through emission adjustments, limiting the overall improvement in agreement with observations. In regions where the sensitivity of NO2 columns to fossil fuel changes is low, the projected prior uncertainty in observation space is small compared to the observation error. As a result, the Kalman gain is reduced and posterior adjustments have a limited impact on improving the fit to TROPOMI, even if emission increments are large in state space.

A full quantitative assessment of this projection would require explicit evaluation of the prior and posterior covariance in observation space, which is beyond the scope of this study but represents an interesting avenue for future work.

- How robust the comparison to the in situ CO2 data can be for the evaluation of the FFCO2 emission estimates if it supports such an increase in national budgets ? Isn't the situation for these stations identical to that for OCO-2, i.e. they mainly catch signal from ecosystems, so that the changes due to corrections to FFCO2 emissions may just slightly compensate for the much larger uncertainties from these ecosystems (see lines 4 and 23-24) ?

We have now commented on the limitation of comparing to these rural in-situ CO2 sites, which are designed to target background CO2 levels and will be dominated by the impact of biogenic rather than fossil fuel fluxes. We also comment on the equal limitations of comparing to urban sites in close proximity to combustion plumes, when model data is sampled across relatively coarse (~25km) grids, systematic model under-estimation occurs. We believe it is interesting to evaluate our posterior estimates against a range of independent datasets, while keeping in mind the limitations of each.

The authors should document the NOx emission estimates from the inversion and the corresponding increments to the CAMS-REG inventory. This should be one of the main results from the study and this would probably help answer some of the questions raised above.

This has now been included in Figure 6a. With individual country time series for NOx prior and posterior in appendix, A2.

Minor comments:

- line 11: the name of the inventory behind the emission ratios deserves to be given here since the quality of their estimate is critical.

The inventory used for the emission ratios was the CAMS-REG v8.1 inventory, and this has now been explicitly stated.

- line 21-24: clarify that these statements are about inversions assimilating CO2 data from current observation networks. "yet it is complicated ..." does not correspond to bottom-up emission monitoring.

We have explicitly stated we are talking about top-down methods.

- line 28: the biogenic and agricultural emissions are probably too high to support the statement "they originate almost exclusively from combustion"

We have changed the wording to 'mostly'.

- line 60: NOx and CO2 emissions

Now included.

- lines 64-65: I do not understand the meaning of "with a reported single measurement uncertainty of 0.7x10^15 molec/cm2"

We agree that the previous wording was unclear. The quoted value refers to the single-measurement precision (random uncertainty) of the NO2 column retrieval under clear-sky conditions. We have revised the text to clarify this and to express the uncertainty in more interpretable terms (percentage range), following the methodology and validation literature.

- line 74-75: why not focusing the evaluation of the inversions against OCO-2 data over these major industrial regions and urban centres ?

We did not perform a targeted urban/industrial evaluation because the coarse model resolution (~25 km) spatially averages emissions, effectively diluting strong point-source signals within grid cells. This limits the sensitivity of the modelled CO2 fields to localised industrial enhancements, making such a focused evaluation less robust with the current setup.

- lines 78-79: why do you need and can you really get a consistency between the qa values used to select the TROPOMI and OCO-2 observations ?

We see that the previous wording could imply a direct consistency between the qa values used for TROPOMI and OCO-2, which is not the case. These quality flags are instrument-specific and are not directly comparable. We have revised the text to clarify that we apply the recommended quality filtering independently to each dataset.

- section 2.1: do the author use TROPOMI and OCO-2 observations over land only ? Figure 1: do the numbers correspond to GEOS-Chem land grid cells only ? (OCO-2 is missing in "one clear-sky TROPOMI observation...") ?

Observations are not restricted to over land only, this analysis uses all grid cells in the domain including ocean. We have added OCO-2 in description of clear sky.

- lines 96-97: "Our simulation..." is unclear

This has been expanded to ‘The forward model simulations used within the inversion framework’.

- lines 134 and 137: NOx and ffCO2 emissions

We clarify that only NOx emission scaling factors are included in the state vector and updated within the EnKF framework. Fossil fuel CO2 emissions are not directly optimised, but are derived diagnostically from the posterior NOx emissions using prescribed emission ratios following the inversion. We have revised the manuscript to make this distinction clearer in the relevant sections.

- line 150: covariance -> variance

Changed

- Section 2.3: the 1-day length of the assimilation window is a prerequisite to understand that there is no indication about the temporal correlations in Pf at lines 152-167.

We have now stated this more explicitly.

- Equation 4: the notation "r" is not very suitable here because it is used for all types of correlations in this manuscript

We have changed to rho.

- Figure 3: what are the « prior ensemble errors »: the variance of these errors for the NOx emissions ? to which extent are they consistent with the theoretical Pf ?

By “prior ensemble errors” we refer to the ensemble standard deviation derived from the sampled prior ensemble. These are generated from the prescribed covariance matrix (Pf) and therefore are consistent with its structure, subject to sampling variability from the finite ensemble size. This has been clarified in the figure caption.

- l177: the use of 1-day assimilation windows implies that the potential temporal correlations of the uncertainties in the prior estimate of the emissions are ignored. It deserves a comment.

This has been included in our discussion of the limitations.

- l202: the word « Europe » is unclear here; it seems that it includes North Africa (l207).

We have clarified this by specifying that the GEOS-Chem European domain includes parts of North Africa.

- Figure 6: why plotting the emissions with a log scale ?

This was originally used to allow a common axis across countries for comparison, but has now been changed to more clearly represent the magnitude of posterior changes.

- Figure 6 and line 230: I do not understand how a 30% change of annual emissions in the Netherlands can correspond to 80% of the prior uncertainty in these annual emissions. The relative prior uncertainties for the daily emissions at the model resolution hardly reach 60% (figure 3) and there is no spatial or temporal correlations in these uncertainties, so they should strongly decrease when aggregated at annual and national scales. Are there missing information which could explain how to bridge this gap ? A bar plot with prior and posterior national and annual budgets with the corresponding uncertainties would be useful.

We thank the reviewer for this comment. When addressing our method, we realised the reviewer is right and we were not properly aggregating the uncertainties for annual budgets assuming random uncorrelated uncertainty in time. We have corrected this, and find that annual emission changes for the 12 countries represent between 6 and 70 times the prior uncertainty. We make comment on the relative magnitude of these changes. Further, the bar chart with prior and posterior national annual budgets and corresponding uncertainties is included in the comparison with EDGAR (Figure 8).

- line231: not really "the largest relative changes", but rather the largest ratios between the relative changes and the prior uncertainties ?

These results have changed (as previous comment). So the description of this figure is updated, and the above wording is no longer used.

- l239-240: unclear; these lines seem to claim that the traditional week days to week end cycle of the emissions would be an artefact in the temporal profiles of the inventories, while the loss of this cycle by the inversion is actually a potential matter of concern.

We have revised the wording in this section to clarify our intent. We do not suggest that the weekday–weekend emission cycle is artificial. In fact, the posterior temporal profiles (e.g. Figure 6a) retain a clear reduction in emissions during weekends, consistent with the prior.

However, a limitation of the prior temporal profiles is the absence of variability within these categories: all weekdays within a given month are prescribed identical emissions, and all weekend days are likewise identical. This results in two distinct branches in Figure A4, corresponding to fixed weekday and weekend emission regimes. In the posterior, these prescribed patterns are relaxed, allowing variability between individual days and weeks. As a result, the two discrete branches merge into a more continuous relationship. This reflects the introduction of day-to-day variability by the inversion, rather than the removal of the underlying weekly cycle.

We have updated the manuscript text to better emphasise this distinction.

- Figure 7: considering the level of correction for the FFCO2 emissions and thus, presumably, for the NOx emissions, the impact of the inversion on the fit to the EEA in situ NO2 observations appears to be weak. However, this is consistent with the weakness of the improvement of the fit to the TROPOMI data. Same comment on Figure 8 (regarding Figure 9: the text already states that the impact of the inversion on the fit to the OCO-2 data is weak).

We have added more discussion on the limitations of comparing with both the largely urban focussed EEA network (which leads to strong negative bias when sampling coarse model grids), as well as the background/rural focussed in situ CO2 (which have less sensitivity to ffCO2 fluxes). The addition of Figure 11 hopefully demonstrates the impact of

inversion on agreement with these networks across the region for each season. Hopefully, it is now more clear that the EEA NO2 network generally shows clear improvement (apart from a few cases in summer), while the in-situ CO2 comparison is more varied.

- l249: the improvement of the fit to the in situ NO2 data is larger in summer, however the smallest corrections to the FFCO2 emissions (and thus, presumably, to the NOx emissions) occur during this season. How to explain this ?

In our manuscript we noted how the improvement in correlations is most noticeable in summer, since data go from having no correlation ($r<0$) in prior to having some level of positive correlation in posterior. The reason for the limited correlations in the summer months is perhaps the impact of shorter NOx lifetime, meaning plume decay is rapid so comparing to model data gridded over relatively large areas leads to higher discrepancy. Additionally, the impact of uncertainty in soil and agricultural NOx fluxes have more impact in the summer. However, the actual magnitude of change in correlation is actually more significant in some of the colder months. Additionally, the MAE and bias improvement are significantly better in winter, when emission corrections are largest. This larger improvement in agreement in winter, followed by autumn is now demonstrated more clearly in Figure 11.

- l251-256: such a discussion questions the use of these in situ NO2 sites to evaluate the inversion results. Should the authors select rural sites only for the evaluation of the inversion ?

Due to the nature of the EEA data, there are limited sites in rural areas. Additionally, there are limitations to both urban sites in close proximity to plumes, and rural/background sites, which are less sensitive to changes to combustion fluxes. We believe it is worthwhile making comparison to both types of in-situ data, while making the limitations of each clear in our discussion of the results.

- Figures 8 and 9: the positive biases in the comparisons to OCO-2 vs. the negative biases in the comparisons to the in situ CO2 data raise concerns. How to explain this ?

The difference in bias reflects the distinct sensitivities of the datasets. OCO-2 measures column-averaged CO2 and is influenced by large-scale transport and boundary conditions, which can introduce a positive bias if background concentrations are overestimated. In contrast, in situ observations sample near-surface concentrations and are more sensitive to local fluxes and boundary layer processes, which can lead to a negative bias if surface emissions or vertical mixing are underestimated. Additionally, the model's lowest vertical layer represents an average over a relatively coarse layer in the lower atmosphere, which can further contribute to a negative bias when compared with near-surface point measurements. As a result, biases can differ in magnitude and sign between the two. We have clarified this in the revised manuscript.

- The maps are generally too small and difficult to read.

Figures have been adjusted to be bigger in size.