Author response – RC1

We thank the reviewer for their thoughtful comments. We respond to them below: the comments are copied hereafter and shown in black, our author responses in blue; suggested new manuscript text is indicated in green. New line numbers in the revised manuscript are provided.

A comparison is presented of three methane (CH4) retrieval products for the TROPOMI instrument. Inversions are performed using the TROPOMI data, the CHIMERE model and a variational data assimilation system. Substantial differences in European fluxes were derived for the inversions using the three TROPOMI retrievals (SRON, BLENDED, WFMD) and these each differed from inversions using the surface data. This result has important implications for the use of TROPOMI data in regional methane inverse modelling.

Overall, I find the study to be important, timely and thorough. However, I feel that the structure of the paper could be improved for clarity and brevity. In particular, I would urge the authors to consider the following:

1. Are all OSSEs strictly necessary? In particular, I wondered if the “diff” OSSEs could be cut (or at least moved to the Appendix/Supplement), without detriment to the paper.
2. In many places, the structure and text could be improved for readability (see suggestions below).

In this study, the OSSEs are the key method to understand the drivers of the differences in emission estimates between inversions. The « *Diff* » OSSE scenarios address the impact of the differences in XCH4 spatial and temporal distributions on retrieved emissions: going further than the random perturbations in other OSSEs, they allow to evaluate the impact of bias between observation datasets. For this reason, we have chosen to keep this part. In the conclusion, we highlight the value of these simulations.

Regarding the structure, we have made changes (in particular in the abstract, Section 2 and the conclusion) to make the text more concise. The structure has been adapted (e.g., the considerations on observation errors in Section 2 were grouped in 2.1.3) and some elements that were not necessary for the understanding of the results have been removed. We hope that these changes will clarify the key statements of the article and improve readability.

General comments

1. I don’t understand why the WFMD product leads to substantially lower fluxes than the other inversions. The terrestrial mole fractions seem to be between those of SRON and BLENDED (Figure 9), and the lateral boundary conditions seem to be similar to BLENDED (D5). The authors address this on L532: “Therefore, the strong negative increments on the Inv-WFMD fluxes result of a complex balance between the local gradients of the increments on the background and on the fluxes: the system could have difficulty separating both when using the WFMD observations.” But can it really be that complicated? If the boundary conditions are roughly similar between two products, but the terrestrial mole fractions are mostly higher for one (Figure 9), then surely, the fluxes for that product must be higher, not lower? Furthermore, it’s not clear why their explanation (that the fluxes and

background can't be easily separated) would only apply to WFMD. I wonder if there could be a bug here…

The lower emission estimates derived from the WFMD inversion are indeed a crucial element of analysis in this study. This element is discussed at the end of Section 3.3. We did not manage to identify a unique origin of these strong negative increments: the distribution of obs-sim is overall positive (Fig. 9) but with values closer to 0 than SRON, so we expect lighter increments than for SRON inversion – this is not the case. We investigated the background optimization (strong negative increments could have compensated strong positive increments on the background, which is not the case) but it was not conclusive either (Fig. D4, D5).

This is likely due to a subtle balance within the inversion process: the differences in spatial distributions of background increments and emission increments with the other products could come from differences in the separation of background and emissions when using WFMD observations. The high number of retrievals could also result in overfitting in the optimization process, and/or residual XCH_4 uncertainties due to less stringent filtering. This topic has to be further explored to allow better understanding of the process.

The results of Rona Thompson presented at EGU 2025 also showed similar relative differences between products (Prior: 27.4 Tg/yr, SRON: 26.4 Tg/yr, BLENDED: 23.7 Tg/yr, WFMD: 19.0 Tg/yr, ICOS: 22.9 Tg/yr). These simulations were made with another model (FLEXPART). It cannot be considered a strict validation, but it suggests that WFMD observations do indicate lower CH4 emissions than the other products in Europe in 2019.

Finally, the structure of the Community Inversion Framework (CIF) makes unlikely the occurrence of a bug, as all the inversions are processed with the exact same code and processing.

Reference: Thompson, R., Schneider, P., and Stebel, K.: Using different TROPOMI XCH4 retrieval products in atmospheric inversions of CH4: a comparison and reconciliation over Europe, EGU General Assembly 2025, Vienna, Austria, 27 Apr–2 May 2025, EGU25-9567, https://doi.org/10.5194/egusphere-egu25-9567, 2025.

2. Throughout, why are posterior flux uncertainties not provided, except for in Figure 9 monthly means?

In the framework of 4D-Var inversions, there is no direct calculation of the posterior uncertainty. It has to be made separately from inversions as such. This is a limitation of our approach, as the comparison of emission estimates between several inversions requires the comparison of uncertainty ranges.

Ensemble methods are often used for this purpose, but with a high computational cost. Following this approach, the uncertainty reduction (independent of the observation vector) can be estimated through OSSEs: computing the standard deviation of priors and the one of posterior emissions across the ensemble samples, the ratio of the two gives the uncertainty reduction. For the total budgets, it is estimated to 78% reduction for SRON and BLENDED, 74% for WFMD and 51% for surface stations. However, we do not use these values for setting error bars to our budget estimates, for two reasons: 1) the poor statistics of this ensemble (4 samples) and 2) the lack of proper uncertainty estimation for the prior as it is not provided with the emission inventories.

Figure 9 contains 1-σ ranges, which are not proper uncertainties: they are deviation of the weekly fluxes used for the monthly average. Clarification has been added in the legend of the figure.

Specific comments

Many statements in the abstract are ill defined, or vague:

L 7-8: it's not stated what these increases or decreases are relative to

The increments are relative to the prior, it has been added in the text.

L 8: "Seasonal emissions are highly correlated across the inversions." I'm not really sure what the point of this sentence is, or what it means. Cut?

The sentence was not clear and therefore removed.

L10: What does it mean that the boundary conditions differ "substantially" for WFMD? Also, is this true? It doesn't seem so from Figure D5.

The sentence was modified, as it was confusely mixing increments on the background and on emissions. The comments on background have been removed of the abstract as they are not essential for the main message of the article.

L10: "Evaluation with independent surface stations shows error reduction for about half of the sites, with BLENDED performing best". I think this means that the residual between the observations and the posterior mole fractions is reduced for about 50% of the monitoring sites. But BLENDED showed a reduction in residual for more sites than the other inversions? More precise language is needed.

This statement was rephrased and quantified to enhance clarity: « Evaluation with independent surface stations shows that the residual between the observations and the posterior concentrations is reduced for 37%, 53% and 47% of the stations, respectively, for SRON, BLENDED and WFMD ».

L11: "However, no product is systematically closer to the emissions estimated when assimilating surface observations". This is too subjective a statement. In any case, to me, the SRON and BLENDED inversions looked very similar to the surface inversion for ~10 months of the year, whereas WFMD differs from the surface inversion most of the time.

This sentence was changed to « No inversion provides a systematically closer match to the spatial distribution of emissions derived from surface observations. » The statement focuses on the spatial distributions of emissions.

L13 and L14: What "errors" and "quality filters" are being referred to here?

These mentions have been removed. We refer to « coverage » and « individual observation error » in the context of OSSE to make the sentence clearer.

Main text

L22: "with higher rates OF INCREASE over the…"

This was corrected, but the sentence has been removed for concision.

L39: "relative", rather than "relatively"

The change has been done.

L45: Shouldn't a range of wavelengths be provided ("spectral range")

The sentence has been removed for concision, the SWIR wavelength range of TROPOMI is provided in Section 2.1.1 (L.96).

L124: Briefly (1 or 2 lines) describe the de-striping procedure

The procedure is now described: « This empirical approach consists in removing the CH4 background by a median smoothing in the cross-track direction, and then computing a per orbit stripe value as a median in the flight direction, which is used for correction».

L137: We keep only THE highest quality…

The change has been done.

Equation 1: define sigma and sigma_hat

Equation 1 and other considerations on the definition of errors have been grouped in Section 2.1.3 (L.195-214). Sigma and sigma_hat are now defined in the text (L.204-207).

L164: Couldn't it be confusing to use Δx here? It could imply only in one horizontal coordinate. Just say within 0.01 degree lat/lon?

The description has been removed for concision and to avoid confusion, as suggested. We only refer to « common » observations, which we think is clear enough.

L180: This is the first reference to a figure (Figure 5). I presume the journal will require that figures are referenced in order?

The reference has been removed to ensure figures are referenced in order.

L208: Why is this interesting? It doesn't actually say, but seems to be implying something. Is this sentence needed?

This sentence was indeed confusing. It has been removed.

L215: "e.g. in Scandinavia". Be more specific: what are the patterns in this region.

The relative comparison of the patterns in Scandinavia has been added: « higher XCH4 than SRON for WFMD and lower for BLENDED, as seen in Figure 2 ».

L216 "The temporal variations… show consistent patterns across the products and align rather well with GOSAT". What is the basis of this statement? To me, GOSAT looks very different to SRON and WFMD, but similar to BLENDED.

The analysis of temporal variations has been improved (L.179-187). Differences were quantified and the comparison with GOSAT has been clarified. In the initial version, authors wanted to compare relative variations, but the « relative » was not clearly mentioned, making the statement confusing. It has been changed in the new version.

L234: Is this really an order of magnitude? Isn't it about a factor of 3?

The change has been done.

L240: I don't understand how the 0-5% difference is quantified to a profile. Is this per level?

It is a per level difference. This has been clarified in the text: « The per level relative difference between SRON/BLENDED and WFMD vertical profiles is below 5% for all levels. »

L325: Do you really mean “surface roughness” here? If so, you could use surface roughness (the meteorological term) as a filter, rather than topography…

The use of « surface roughness » has been changed to « subpixel topographic variability ».

L390: If the dataset is split randomly, isn’t there going to be substantial correlation between the testing and training sets that influences the metrics? Most of the testing set will be adjacent to points that have been used in the training. It would be preferable to have the testing and training set be separated in space/time (or some other factor).

The random split of the dataset can indeed generate correlation between the training and testing sets. A sensitivity test was run using a different split for the SRON-WFMD pair, with Jan. to Oct. 2019 as the training dataset and Nov./Dec. 20219 as the testing set. The performance of the prediction was slightly lower, with RMSE of 8.9 ppb (in comparison to 8.1 ppb with the random split) and R^2 of 0.51 instead of 0.58.

However, we are not directly interested in the performances of the model but rather in the feature contributions in the calculation of the prediction. Indeed, the SHAP value analysis was very similar in both cases, with differences of average |SHAP| (Fig. 7) between both cases inferior to 0.1 ppb. To avoid seasonal bias that could be introduced by the temporal split of the data, we chose to keep the random split of training/testing set.

Figure 7 caption: this figure needs explaining more thoroughly

Explanations have been added to facilitate the reading of the plot.

L413: since this is the first line of a paragraph, restate this initial sentence so that it reminds the reader what you’re talking about.

The reminder has been added.

L450 and elsewhere: try to avoid subjective terms like “best”.

The change has been done here and for other occurences of such subjective words.

L472: You can’t say that the inversion “correctly” fits the data as there will always be errors. Just say that the residual is reduced after the inversion (which it must be, if the inversion is working correctly).

The sentence has been restated accordingly: « the residual (difference between the observed XCH4 and the posterior simulations) is reduced after the inversion ».

Figure 10: why is this the only place where emissions uncertainties are provided?

See answer to General comment #2.

L506: “Yet, the amplitude of the emission peak raises questions about its origin.” This sentence implies something but doesn’t spell it out. What are the questions, what could be the origin? Or do you just mean that you don’t believe that this peak is real? If so, say so, and justify your reasoning.

The sentence has been changed to «Yet, the origin of this emission peak has not been clarified.». We investigate the origin of the peak but could not find a process that fully explains it. The elements cited just before, the slight decline in observed XCH4 and the increase in the lateral boundary conditions in April, are remarkable elements but do not provide a clear explanation. The-

refore, to avoid unclear interpretations, we only mention that we cannot certify what is causing this peak.

L507, L512 and elsewhere, don't start paragraphs with "However,", "Moreover", etc. Each paragraph should tackle a single idea. These words imply a continuation of an idea.

The change has been done here (L.468, 473) and for other occurences (L.526, 560).

L556: Avoid "better"

The change has been done here (L.517) and for other occurences of such subjective words.

L602: "It can appear in the optimization process in variational inversions." This is a sentence fragment. Reword.

This sentence is not necessary for the explanation, it has been removed.

L606 – 629: Could this section be removed?

See answer to General comments.

L635 – 637: I don't know what this sentence means. Rewording is needed for clarity.

This sentence has been rephrased (L.595). Here we compare the difference between observed and simulated XCH_4 based on: 1) the prior emissions 2) the posterior emissions estimated for the inversions of each TROPOMI product.

For SRON and WFMD, the average difference obs-sim and RMSE are higher (in absolute values) in comparison to the prior ones. For BLENDED, they are lower (in absolute values).

Conclusions: I think the conclusions are far too long and contain several statements that aren't really justified, given the results. The authors should refocus this section for concision.

The conclusion has been the object of a concision and rephrasing work, to refocus the main message of the article and to avoid making confusing statements that are not well supported by the results.

L661 – 662: This one-sentence paragraph doesn't seem necessary for a conclusions section. It doesn't really say anything.

This sentence has been removed for concision.

L675-676: What is meant by a "proper formula"? Wouldn't it be better to say what is physically needed here. I.e., how should the error be appropriately calculated.

We have described more thoroughly the calculation of the error that we recommend (L.641-644): we suggest a linear regression of the scatter of the observations relative to TCCON to derive the individual observational error, similar to what is done for the WFMD product (Eq. 1).

L678: Are model biases really a key outcome of this study? Do we need to state this here?

The model error was not directly studied. This study identified issues that could be related to issues in the transport modeling part of the simulation (reversed seasonal cycle of CH4 emissions, peak of emissions in April/May...). But there was no thorough analysis of them, they are just some leads to interpret the results. Therefore, we have removed this sentence to avoid confusion.

L679: “Refinements in the configuration of the inversion system are thus also essential to enhance the consistency and robustness of emission estimates.” What does this mean? Why is it essential? What about your results indicates this to be true? (If you are just saying that inverse modelling systems need to be improved, you can cut this, as it’s well understood, and you don’t address this in your paper).

Indeed, improvements of transport and inverse modelling were not addressed in this study. The sentence has been removed.

L693 – 694: Your results don’t show why we need higher resolution, and, clearly, global inversions would be preferable. You can safely cut these lines.

The sentence has been removed.

L697: Does your work really suggest that joint in situ and TROPOMI inversions would improve matters? At the moment, you show major differences that are difficult to reconcile. Perhaps, if we knew how to appropriately reconcile these systematic differences. But at the moment, I wonder if your work shows that in fact, we’re not ready yet?

Sections 3.3 and 3.5 highlight differences between results of the in situ and satellite inversions. Reconciling the two types of inversions is a great perspective, but indeed this work indicates that it is still a challenging problem. The sentence is more of a long-term perspective than a conclusion of this study, we have removed it from this article.

Supplementary material

Figure D4: Is this missing a 4th row (mentioned in the caption)?

The caption has been updated. The 4th row existed in a previous version, but was removed later. Also, this figure is now D5, to ensure the correct order of figure referencing.