Author response v2 - 10/10/2025

Response to 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 CH_4 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 XCH_4 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. Therefore, 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.

Response to RC2

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.

This manuscript by Sicsik-Paré et al. presents a comparative analysis of three XCH_8 products derived from TROPOMI observations—SRON, BLENDED, and WFMD. The authors examine various parameters influencing discrepancies among the XCH_8 datasets, highlighting aerosol scattering and SWIR albedo as primary contributors. Additionally, the study utilizes these products in inverse modelling with the CHIMERE model and a data assimilation system to estimate surface methane fluxes.

Overall, the manuscript provides valuable insights for the atmospheric methane community, especially in the context of inverse modelling and satellite product evaluation. It demonstrates the potential and limitations of different XCH_8 retrievals and their influence on flux estimates. I recommend publication in ACP after addressing the following general and specific comments.

General comments:

1. The manuscript is generally well-structured but somewhat lengthy in certain sections. For instance, the detailed description of the CHIMERE model in Section 2 could be moved to the appendix to improve the flow and readability of the main text.

The feedbacks from the different reviewers all suggested to improve concision, especially in Section 2. This section was reorganized to remove unnecessary statements that did not support the main message of the article. In particular, we have made the description of CHIMERE more concise. However, we prefer to keep this subsection in the main text, as the description of the model set-up is usually part of the main text for inversion articles.

2. The WFMD product reports the highest absolute XCH_8 values among the three datasets and shows the largest bias relative to TCCON (7.5 ppb; see Table C2). While the WFMD XCH_8 dataset shows moderate differences from both prior and posterior simulations compared to SRON and BLENDED, the fluxes derived from WFMD are consistently the lowest (Figure 10b). This discrepancy raises discussion here.

Could the lower flux estimates be a result of overestimated background subtraction? Furthermore, WFMD includes approximately 30% more retrievals than the other products. Could this may be due to less stringently filtered, allowing residual uncertainties to remain, which could contribute to the observed differences in flux estimates?

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 from 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.

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.

3. While the study uses surface observations for the inversion, have the authors considered assimilating ground-based total column data, such as from TCCON or COCCON? These measurements are widely used to validate satellite products, as both observe total atmospheric columns and are therefore directly comparable. Incorporating such datasets could provide a more representative constraint on regional methane emissions and may help improve flux estimates, particularly in regions where surface observations are sparse or not fully representative.

The use of TCCON data in this study is limited to the comparison of XCH_4 datasets and is presented in Annex C. We do not use the data for the comparison of the emissions estimated from the inversions, thus it is not used as a validation dataset. We use in-situ ground-based measurements for the validation (Sections 3.3, 3.5).

We have considered the assimilation of ground-based total columns from TCCON or COCCON. We decided to assimilate in situ measurements as an independent dataset because the 19 stations used in this study offer higher spatial coverage than the 5 (COCCON) + 7 (TCCON) sites available in 2019 (with 2 redundant sites in Karlsruhe and Sodankylä).

Specific and technical comments:

Line 126-127: the authors reference previous work to justify scaling the errors by a factor of 2, but a brief explanation here would enhance clarity for readers unfamiliar with that context.

Considerations on the definition of errors have been grouped in Section 2.1.3 (L.195-214). Specifically, we have described more thoroughly the error in the SRON product, and explain that the product authors « suggest to multiply them with a factor 2 to reflect the scatter of errors in the TCCON validation ». This is based on the ReadMe of the SRON product (see reference to Landgraf et al., 2024, page 10 of the document), but no further explanation is available.

Line 207: Figure 3 is referenced before Figure 2 in the text. The order of figures should be revised for logical flow.

The order of references has been reversed in the text (L. 175,176).

Line 207-208: The phrase "BLENDED lower concentrations" is unclear. Does this mean that only the lower end of BLENDED values agrees well with GOSAT? Consider clarifying this point. Additionally, including GOSAT's spatial distribution or mean value in Figure 3a would improve visualization and support the comparison.

The explanations have been clarified: the statement that BLENDED XCH_4 are lower than SRON and WFMD for all months is quantified (L. 184,185), and the statement that BLENDED is the closest to GOSAT is mentioned (« Only BLENDED aligns rather well with GOSAT, due to its correction of the TROPOMI-GOSAT bias »).

We chose not to compare averages of TROPOMI XCH4 with GOSAT XCH4 for the readability of the Figure. Also, the sampling is very different for both instruments: the comparison of averages can therefore deceive the reader in thinking than the differences are due to biases between products, whereas they can be caused by differences in coverage.

Figure 3: It seems the errors for SRON and BLENDED largely overlap in panels (b) and (c). If this is the case, please clarify it in the figure caption.

BLENDED errors are retrieved from SRON, so errors from both datasets are exactly equal. It is explained in the main text at L.129 and 196. The sentence « SRON and BLENDED errors are exactly similar, as BLENDED errors are directly retrieved from the SRON product. » has been added to the caption of Fig. 3 to improve readability.

Line 240: typical of SWIR retrievals >>> type?

This has been changed to « characteristic of SWIR retrievals ».

Line 408: The analysis indicates that differences are closely linked to across-track pixel index, aerosols, and SWIR albedo. While the latter two are physically intuitive, the role of across-track pixel index is less clear, as it represents only a positional identifier. Could these index be indirectly capturing information related to albedo, aerosol, or stripping artifacts in the retrievals? Further clarification would be beneficial.

The link between the positional across-track ID and striping patterns is described in L.342. A reminder « across-track pixel index (thus striping patterns) » has been added at the mentioned location (now L.365) to remind the reader of the link between the ID and striping effects.

Line 519: Background removal is known to be challenging. The manuscript would benefit from a brief explanation of how the background is defined and removed in this study.

The background mentioned here is composed of the lateral and top boundary conditions, the initial conditions and the stratosphere concentration field that are taken from the CAMS global product. It is defined in Section 2.4.2 and we have added here a reference to this section.

Line 531-532: result of >>> result from?

The correction has been done.

Line 560: The authors note poor improvement over Scandinavia. Could extending the study period improve coverage and reduce uncertainty through increased data availability? This point deserves further discussion.

Poor improvement over Scandinavia is due to the lower coverage in this region (Figure 1), in particular in winter (many pixels with no observations for JFM and OND in Figure D3).

The extension of the study has been considered to generalize the analyses and capture seasonal signals (only one year is not sufficient, some temporal variations could be specific to this year). It has not been implemented because of the already high number of simulations performed in this study, and their high computation cost. However, it is not clear that extending the study period would reduce uncertainty: the increased coverage would provide more data, but in the meantime the inversion would have more emission and background components to constrain, resulting in the same « constraint capacity » of the inversion. The ratio between the number of observations in Scandinavia and the one in other parts of the domain would remain approximately similar, thus the inversion would behave similarly and provide low constraints in Scandinavia.

The study of high latitude regions has to be the object of dedicated studies, which has already been the case using TROPOMI data: the studies of Tsuruta et al. (2023) and Lindqvist et al. (2024) are cited in the manuscript, see also the more recent study of Kivimäki et al. (2025).

Reference : Kivimäki, E., Aalto, T., Buchwitz, M., Luojus, K., Pulliainen, J., Rautiainen, K., Schneising, O., Sundström, A.-M., Tamminen, J., Tsuruta, A., and Lindqvist, H.: Environmental drivers constraining the seasonal variability of satellite-observed methane at Northern high latitudes, EGUsphere [preprint], https://doi.org/10.5194/egusphere-2025-249, 2025.

Appendix C: COCCON is increasingly used as a complementary dataset to TCCON, particularly in regions lacking TCCON coverage. The authors could also consider incorporating COCCON measurements to provide additional reference data for satellite validation, thereby enhancing the spatial representativeness of the evaluation.

See the answer to General comment #3: we agree that COCCON would provide additional data enhancing the comparison of TROPOMI XCH_4 products. The use we make of TCCON data in this study is limited to the comparison of XCH_4 datasets: as it is not directly used for the comparison of the emissions estimated from the inversions, it is not an evaluation of our results but more a complementary comparative description of XCH_4 products. For this reason, the section was moved to Appendix and we did not consider using more data from COCCON. For further studies, it will be very beneficial to use both observational networks to enhance the spatial representativeness of the analyses.

Response to RC3

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.

General remarks:

At the beginning I would like to underline that this article is definitely worth publishing as its message is of the most importance while satellite products like CH4 concentration distribution are becoming more and more popular and conclusions taken from those measurements have political and economic importance. Authors did extremely good job and provided variety of analyses regarding three basic products related to CH4 column concentrations. The main aim of this article is to answer the title question. The role of scientific article is to give evidences and logical arguments with statistical approach to set up the thesis and confirm it (or discard it).

The scientific level of this article is very high. Large part of the paper describes methodology applied to obtain particular products from TROPOMI instrument. Secondly, modelling of the methane fluxes was described. This was done to bring the final product, i.e. CH4 fluxes to the readers. From the point of view of scientist not working with satellites measurements the paper is very long and frequently refers to other publication describing in detail particular techniques and methodologies. In my opinion it resulted with a patchwork of information not deep enough to understand the products without reading those referred publications and making the storyline heavily disturbed. All this looks like authors wrote the paper and gave the title question at the end. I bet that readers of this article will be curious what is the answer to the question. And presentation of that problem should be a main goal of this paper. Or authors can change the title easily to definitely more scientific oriented, like “comparison of three products…..”, where such structure of article will be consistent with the title.

This comment and some of the comments below are driven by the fact that the title of the manuscript was misleading. The title question indeed suggested that the question will be answered in the article, and we provide more of a comparison, rather than we provide an exhaustive answer to the question of the consistency of emissions from the inversions. Here, we compare the products and the emissions estimated from inversions, and try to investigate the drivers of the differences that we see.

The title is revised to “Assessment of the differences in European CH4 emission estimates from three TROPOMI products”. With this title, the structure of the article is more logical: we first compare directly the products (XCH4, coverage, errors, vertical profiles) in Section 2, then look for the drivers of these differences in Section 3.1 and continue the comparison in the next sections: obs-sim distributions in Section 3.2, emissions derived from inversions in Section 3.3, evaluation with surface stations in Section 3.5, with another look to the drivers of differences (for emissions this time) in Section 3.4. We investigate when the products are consistent or not, but do not provide a definitive answer to the previous title question, which has been changed.

The consistence of flux/emission product should be thoroughly analysed with detailed prepared tools. This is where OSSEs helps a lot and article clearly benefits from this approach. Consistence is also expressed in statistical way, and here the maps are extremely helpful. However to answer the question authors should present a clear and trace line leading from methodology to consistency at pixel, region, country and continental scale. Here, analysis is dissolved in many technical references which are not relevant to the final answer.

We have modified the manuscript to clarify the organization of Section 2 (see answers to specific comments), make Sections 1 and 2 more concise. We have also tried to highlight a clearer structure, introducing it more carefully in the abstract and introduction, and bridging with the conclusion. Content that was presented in Section 2 and not used in Section 3 has been removed, and we have removed the subjective statements.

I see clear lack of hypothesis verification and many subjective statements not explaining if we can or cant obtain clear consistency of 3 TROPOMI products. My expectation for this article would be a detailed discussion what affects consistency of the 3 outputs in terms of methane flux distribution (temporal and special).

Discussion of similarities and differences between the products. Those only which are relevant to final conclusions.

Discussion of metrics applied to judge consistency of 3 products or any pair of those.

Discussion of results in form of the prove of the hypothesis with detailed description of the discrepancies and similarities of the product results (final ch4 fluxes)

Discussion of possible errors and uncertainties of the fluxes and its relation to consistency.

If authors find it necessary comparison with ground-based observations which in my opinion is irrelevant to the question (in title).

One would expect that the test of consistency can be done using known emission distributions for methane (large point source and large area sources). There are some verified sources of methane which can be potentially used for such purpose. However, again, agreement with real data is not matter of product consistency but rather reliability. Authors should look at the comparison of the products and condition when they deviate from consistency. That would be an answer to the title. In the form they proceeded it's a mess.

4D-Var inversions have been performed in this study, and therefore we do not have direct access to posterior uncertainties. This lack of uncertainty range intrinsically limits the analysis of consistency, as we cannot statistically check if estimated emissions are consistent or not. For this reason and to clarify the structure, we have modified the title of the article.

However, the manuscript is a discussion of similarities and differences between the products, at different steps of the inversion process: observations, comparison simulations vs. observations, derived emissions. At every steps, the distributions are compared in terms of average, deviations, spatial and temporal variability. When differences are observed, they are investigated (Sections 3.1, 3.4) through detailed analysis with metrics : SHAP values to evaluate feature contributions to XCH_4 differences in Section 3.1, relative increments in the case of OSSEs in Section 3.4.

The objective of this study is to assess whether TROPOMI data users would find estimated emissions that are consistent if they perform inversions. In this case, we evaluate « consistency » with respect to the aim of building emission inventories at the pixel, country and continental scale. In this perspective, results can be compared to the ranges of estimates of top-down budgets: 31 [24–36] TgCH4/yr for Europe (not exactly the same mask as in this study) for 2010-2019 in Saunois et al. (2025), and 19 [15–23] TgCH4/yr for EU27+UK for 2009-2021 in Petrescu et al. (2024), cited in the manuscript. We have clarified the way we consider the consistency in the conclusion, but did not include the comparison of the ranges as the differences in covered areas and periods could be misleading for the reader.

Also, 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.

Specific comments:

Abstract.

Abstract should be a kind of teaser for what reader will find inside the article. It is. However, the expectation of answering the question is not confirmed in it. One can have feeling that the article will contain a chaos of information what indeed is true.

The obvious comment that XCH4 differences are related to aerosol and albedo one can find in any of publication related to any of those products. Authors should be a bit more constructive here giving something new in that field (e.g. quantification of those effects).

We quantified the impact of these parameters on the XCH_4 differences between pairs of products. We have evaluated the impact of XCH_4 differences on the emissions through the *Diff* OSSE scenarios, but did not performed separate analysis for albedo and aerosols. It would have required to assimilate bias-corrected products in inversions, and thus modify the XCH4 distributions. This is an interesting perspective and we mention it as future work in the Conclusion.

However, it was chosen not to perform this analysis in this study for three reasons. First, there is already a lot of content in the article. Secondly, the only difference between SRON and BLENDED is the XCH_4 bias correction of the effect of albedo, aerosols, striping, etc. in BLENDED. The small differences in the results of inversions suggest that these parameters do not impact that much the derived emissions (this statement is qualitative, its quantification is required in future work). Finally, we have chosen to focus on parameters that do not require to modify the XCH_4 distributions, in order to assimilate products as close as any data user would do, with the perspective of application uses. We agree on the fact that the quantificationof the albedo and aerosol impact on the emissions is an important next step.

Comprehensive abstract should not contain contradictory statements : e.g.

"European CH4 emission budgets show relative increase of 2%, -1%, 33%..." - whatever relative increase is, its clear products are not consistent. Or they are, because 33% in compare to 2% is not much?

"Seasonal emissions are highly correlated…" – so, probably we will look for bias in products. So, if the correlation is a measure of consistency the answer is yes, if bias – the answer is no.

"Using consistent error definition….for more consistent emission estimates" – what is consistent error definition at this stage its hard to understand but it suggest that it will play a key role in the answer. But it is only "….paving the way…" so most probably the answer is again: No.

Article will be much more understandable if authors will focus on main message which is the answer to the title question and will build the well structured hypothesis verification.

See the answer to general comments regarding the title. The abstract has be modified accordingly to clarify the comparison and make the trace line of the article clearer.

Introduction.

First few sentences (lines 17 – 24) are so much obvious and frequently repeated. Its now primary school level knowledge. I think this article is not giving any new concept about those topics and as being submitted to ACP, every person who will read it will be aware of those basic facts about methane.

Next few sentences (lines 25 – 29) are referring to BU, which is later not compared with nor used in any form. Authors also do not provide any input to this method of methane budgeting. Critical approach to the BU EU methane budget is not relevant if there is no better proposition from authors side.

We agree with the reviewer that there is no contribution to these topics from this study. The first two paragraphs have been grouped and shortened. A few sentences were kept to introduce notations and provide context for why it is important to have robust TD budgets.

Definitely there is a lack of introduction about satellites, especially those used for XCH4 measurements. There is a whole constellation of such platforms nowadays and role of S5P in that landscape would make a good background for this paper. Please, describe or list all products currently used for XCH4 retrieval. This will help to understand the problem of coherence of product.

A more detailed list of the currently used instruments for XCH_4 retrievals has been added to the introduction (L.28-33). The role of S5P is detailed when mentioning the resolution and coverage, as well as with the use of TROPOMI data for various types of emission quantification.

Secondly, please discuss where the products were applied (examples of application of those products – like lines 40 and 41, but with deeper explanation of problem in term of coherence of those products). How much those products differs for O&G facilities, mines, landfills…

The use of several products in a single study is rather limited and described in L.58-64. The of the consistency of the products is not analysed in the studies listed for O&G facilities, mines, landfills, etc. This is what we want to adress in this study.

Be specific while mentioning some aspects of work:

Line 48 – prior vertical profile of what parameters?

The mention « prior profile of CH_4 mixing ratios » has been added.

Line 49 – stratospheric background of what?, aeroslos… - do not leave "…" here. Discussion of those effects are essential for this paper and should be introduced carefully in that chapter (instead of GHG effect).

The « ... » has been removed and background has been described. The mentioned effects are identified through the ML analysis as drivers of the XCH_4 differences between products and are detailed in Section 3.1, but they are not studied in inversions (see previous answer to the comment in the abstract).

Line 49 – what retrieval?

The formulation was not clear, it has been changed.

Line 50 – scattering of what? What spectral range? What issues are mentioned here?

The sentence has been changed to « ... scattering due to the presence of aerosols ».

Line 52 - …routinely updated. – provide the reference.

To be more accurate, « routinely » was changed to « iteratively ». The iterative versions are described in the provided references.

Line 54 – avoid slang, like “beta research product”.

The change has been done, « beta » has been removed.

Line 54 – At this stage destriping algorithm might be at least briefly explained.

A brief description of the principle of this algorithm has been added: « a destriping algorithm, based on moving median smoothing in the across-track directions and flight directions, has been developed for new XCH4 data ».

Line 67 - …the consistency and applicability… those terms require the measure (scale) would be nice to have few words in introduction about it.

This sentence has been removed for concision.

Please explain why Europe is a targeted area for this article and why 2019 was chosen. It is important because later authors discuss a seasonal effect, what in relation to single year is not usual. Especially that there is a definitely longer period available now (2025). Effects in that year might not be representative for a “usual” seasonal singals.

As mentioned at L.69-70, the study focuses on Europe due to its extensive network of surface stations, as we use them to support the comparison. Other regions (e.g., North America) could have been chosen, but we wanted to build on the studies focusing on Europe that have been published in the last decade.

We chose to focus on 2019 to have a full year available (no TROPOMI data before april 2018) and to avoid patterns of emissions related to the lockdowns (2020 and 2021). The extension of the study has been considered to generalize the analyses and capture seasonal signals. It has not yet been implemented because of the high computation cost of the performed simulations. A follow-up study should inquire the long term implications of using a given product rather than the other.

Data and methods

Here, my opinion is that authors can easily shorten this long chapter giving a list of good references to each product. Only aspects important for a latter discussions should remain.

This section has been restructured and many subsections have been shortened for concision, and to ensure we focus on elements that are useful for the analyses of Section 3.

Lines 100 – 108 – are those details important for the products coherence?

This paragraph recalls a few parameters of the TROPOMI instrument and the error requirements, and also serves as an introduction to the description of products. Some unnecessary details have been cut for concision.

Line 115 – This regularisation is more frequent referred as Tikhonov, or Tikhonov-Philips. Again level of description is strange. We cant see any reference to this algorithm anywhere later. Line 116 - …state vector… not explained what it is and how it may affect coherence? Line 119 – VIIRS instrument is certainly important but does use of it influence the coherence of the products?

These sentences have been removed for concision, and because they do not provide useful information for our analyses.

Line 123 – …(qa_value… later in line 137, and variable xch4…in line 152) is such variable need to be mentioned here? If yes please describe what it is.

For the 3 products, the names of the dedicated variables were removed, we use « quality flag » (L.108, 125, 134).

Line 124 – why do we need to know that there is 4731034 observations – how this number affects the conclusions? Table 1. contains this information.

To avoid repetition with Table 1, this sentence was removed.

Line 124 – if destriping is applied to only “new” data and it is important for coherence in 2019 – why authors didn’t insist to apply it also for 2019?

We identified striping as one of the main drivers of XCH_4 differences. However, the destriped data is not available in 2019. Data could have been reprocessed, but at the cost of modifying XCH_4 data. In the perspective of using data « as an user would find », we decided not to reprocess the XCH_4 observations with an additional algorithm.

Line 126 – what is the single sounding precision?
My comment to authors regarding uncertainty calculations: It is a crucial part of this paper and authors underline it the conclusions. Why it is so poorly explained in the chapter 2. Please devote some space to explain directly how you apply the uncertainty chain in your calculations. Give a table with some values.

Considerations on the definition of errors have been grouped in Section 2.1.3 (L.195-214), to make the structure of the article clearer. The terms such as « single sounding precision », « pseudo noise »... have been changed to clearer words. We have summed up the definitions of errors provided in the ATBD of each product (SRON/BLENDED on the one hand, WFMD on the other hand). In both products, the noise in the measured radiances is propagated in the retrieval algorithm. Then, the post-processing of this propagated error is based on the comparison of XCH_4 observations to TCCON : to reflect the scatter of errors in this TCCON validation, different post-processing are implemented across products.

- For SRON/BLENDED, the authors suggest to rescale the provided error by a factor of 2.
- For WFMD, the provided error already take into account a linear rescaling (Eq. 1).

The difference in the individual errors provided by the products is what we want to highlight here.

Line 133 – TCCON observations bias regarding GOSAT not TROPOMI product is not referenced. Explain why do we need this correction and how does it affects the consistency of products?

The removal of this bias is made by Parker et al. (2020), which provide the product. It is applied to the GOSAT observations used to create the BLENDED product (Balasus et al., 2023). The effect on the consistency of products is described further (Section 2.1.3), when comparing the XCH_4 distributions.

Wouldn’t be good if TCCON is used for product consistency check?

A comparison of TROPOMI XCH_4 products to TCCON is provided in Annex C. Initially in the main body of the text, it was moved to Annex as it is not directly used for the comparison of emissions from the inversions. It provides a comparison to independent data, interesting when comparing the products, but does not serve directly the main message of the article.

Line 136 – limitations are extremely important for this discussion – no references, no comments?

These limitations are discussed later (Section 3.1), so we do not insist on them at this specific location of the manuscript.

Line 137 – qa-value – what product it refers to ? SRON?

See the answer to the previous comment on quality flags variables.

Line 137 - …recommended by the authors. what is rationale behind that? Authors of what? Line 141 - …provided errors… please give a reference and describe more specifically. Line 158 – Section 2.1.3. contains very poor description referring partly to section 2.1.1.! And defines as st dev of retrieval noise – not explaining what it is and how does it is calculated. In terms of XCH4 values. Please, provide this information somewhere. Also pseudo noise should be defined if it is relevant to final conclusions.

See the answer to the previous comment on errors: the description of the error definition for each product has been restructured in Section 2.1.3.

Equation 1 – no variable symbol definition is provided.

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.205 and 206).

Line 169 - …is relatively mature product, even though sparse. - If it is important for BLENDED product please indicate whet does it mean.

This subjective sentence was removed for concision.

Line 174 – 9.2 ppb bias is repeated ?

This information is mentioned twice: one in the BLENDED description when mentionning GOSAT, and one in the GOSAT description when making the link with BLENDED.

Line 194 – what scene description variables are important here?

The most important ones have been added (cloud fraction and SZA), the other ones can be found in the added reference (ReadMe of the SRON product).

Line 195 – Spatial patterns of differences…. - Of what?

The sentence has been modified to « of coverage differences ».

Line 199 - …change of pixel size. – why this year was chosen not e.g 2020?

2019 was chosen to check if the change of pixel size has a major impact on coverage (which is not the case), and to avoid emission patterns related to the various lockdowns in Europe (2020 and 2021).

Line 203 – If CFC was exceptional, maybe longer period of comparison (2019 – 2024) would be beneficial to cancel such effect.

The extension of the study period was considered but not implemented due to its high computational cost. Moreover, the variation of coverage due to high CFC is not detrimental for the comparison of emission estimates. Further studies will apply multi-year inversions to assess this impact, beyond the scope of the present paper.

Line 205 – where is a Figure S1?

Figure S1 is available in the Supplementary material. A mention « in the Supplement » has been added for clarification.

Line 206 – there are statistical tests for non parametric hypotheses? What does "approximately" mean here?

This subjective statement was removed, as the gaussian aspect (or not) of the distribution is not used later in the analysis.

Line 209 – 9.2 ppb mention 3rd time

This repetition was removed.

Line 210 – why there is no such comparison at the Figure? Its not clear how the GOSAT average is calculated if there is so big difference between the pixel size and number of data?

The comparison to GOSAT averages was removed, as the large differences in coverage between TROPOMI and GOSAT products makes the comparison of averages not relevant.

Line 213 – There are over statistics describing distributions not only average (expected value).
Line 220 – how does offset is calculated (difference of averaged monthly values?), why not calculate it for each products pair.
Line 220 – what does approximately means - why not to give st dev?

The differences of averaged monthly values have been quantified (average and standard deviation) for each pair of products (L.184-185), and the word « offset » was removed as it was confusing.

Line 230 – what noise authors refer to?
Line 232 – what is "pseudo noise" ?
Line 234 – 3.9 is not order of magnitude lower than 12.2
Line 234 – scaling up the "error" is unusual procedure that requires additional clear evaluation – please provide the reasons behind this transformation. Being smaller is not one.
Line 235 – Author should explain why Eq1 is applied and where does those coefficients come from.
Line 236 – no definition of error is given (anywhere), how can be error understood here : uncertainty, systematic error (bias),
Line 237 – also here observational error is without definition…please provide.

See previous answers to comments on the error definition: we have grouped all the error considerations of Section 2 in L.195-214. We replaced « order of magnitude » by « factor 3 ». We provide the explanation of the error scaling in WFMD (and the corresponding references), and why we apply a similar scaling to SRON/BLENDED errors.

Line 244 – Huge difference of XCH4 (200ppb) is hardy evaluated by fig 4. Can authors provide a panels with a view on higher pressure levels so this, most important part of profile is clearly visible?

Figure 4 has been modified to provide a zoomed view of the high-pressure levels.

Line 275 – please give a reference (…less than 2 weeks)

A reference to a study on the transport of air masses in Europe (Nygård et al., 2023) has been added to justify this statement.

A whole chapter 2.3. should be target at application of products, its importance for emission estimation and coherence of the emission patterns. Details can be linked to literature and equation 2 is not necessary, or authors should refer latter to it discussing the sensitivity of model to pressure profile in particular products if different.

Chapter 2.3 was shortened for concision, Equation 2 was removed as it is not refered to later.

2.4. chapter is again basing on many references and construction of inversion model principle is not relevant if not used later in results chapter. This article is not basing on particular matrices used in eq.3. If statement in line 231 can be explained without introduction of the Eq 3 than authors may revise need of that equation and chapter 2.4.1 and large part of 2.4.2.

Sections 24.1 and 2.4.2 were restructured to make them more concise and remove unnecessary elements. However, we believe that it is worth quickly restating the principle of our inverse simulations, as well as describing thoroughly the set-up to ensure replicability and clear statement of the hypotheses.

Line 305 and 306 – why usually 20 – 30 , what is the cost function , is it relevant for this discussion?

This confusing statement has been removed. The idea was to indicate that the 95% reduction criteria is in practice reached after 20 to 30 iterations. But this is not useful for further analyses.

Line 315 - is it RSD? Why set it to 100% ? How does it correspond to “error” definition in model?
Line 318 – why 2% not 5%

The mention « relative » has been added for the standard deviation. The values have been chosen based on commonly shared values, in particular from the intercomparison of CH_4 inversions in Europe (Ioannidis et al., 2025). The study of Szénasi et al. (2021) also helped chosing these values. The background error was decreased from 10% to 2% to ensure that the weights of corrections on emissions and background are not too unbalanced.

Reference: Ioannidis, E., Meesters, A., Steiner, M., Brunner, D., Reum, F., Pison, I., Berchet, A., Thompson, R., Sollum, E., Koch, F.-T., Gerbig, C., Wang, F., Maksyutov, S., Tsuruta, A., Tenkanen, M., Aalto, T., Monteil, G., Lin, H., Ren, G., Scholze, M., and Houweling, S.: An inter-comparison of inverse models for estimating European CH4 emissions, Earth Syst. Sci. Data Discuss. [preprint], https://doi.org/10.5194/essd-2025-235, in review, 2025.

Line 319 – What does “large” means in reference to model error?

The criterion used for defining « large » differences is > 100 ppb. This is an arbitrary choice to avoid the inversion to focus on very large obs-sim differences.

Line 325 – what is “poor” quantitatively? How does it affect XCH4?

This subjective statement has been removed.

Line 331 – Confirmation is not supported by logic and statistics, model can run wrongly in some areas and this will also result as same distribution of removed observations in all products.

This statement was indeed not well supported, it has been removed.

Results

I personally found this chapter very interesting and well prepared. It contains a lot of earlier briefly or poorly explained information and references. It looks like it was written by different person or group than chapters 1 and 2. My only general comment in relation to the results is to focus on coherence of products (like fig.7). Give as much quantitative comparisons and avoid subjective statements like “could be related” , “possibly” etc.

We have taken this comment into account and have tried to rephrase as much as possible these expressions, while providing quantifications. However, some occurences have been kept to insist on the fact that we suggest a explanation, but cannot certify a unique cause.

Chapter 3.3 - Seasonal cycle and annual budget are the parameters which are affected by boundary conditions and exceptional situations if they are referred to a single year period. In my opinion authors could make the analyses longer for this comparison.

The expression « seasonal cycle » has been changed to « temporal variations »: as mentioned, seasonal cycle cannot be analysed through a 1-year analysis. See previous comments on the extension of the study period on this aspect.

Line 506 – Figure D5 is appearing before Fig D4 is mentioned (line 520)

The order of these figures has been changed.

Line 612 – There is no information what is uncertainty of calculated yearly emissions and it is hard to judge if change from 25.2 to 25.3 is statistically significant. Definitively term “slight” is not quantitative.

See the answer to the general comment for uncertainty considerations. The subjective word « slight » has been removed.

Conclusions

I will repeat my general remark here. I would like to know the answer to the title question. If it hasn't been given in chapter 3, I expected it to appear here.

See answer to the general comment.

Line 654 – what does the “holds great promise” refers to – list a reference or be more specific.

This expression has been removed and replaced by a reference to the review of current and scheduled satellite observations from Jacob et al., 2022.

Line 658 – Why consistency of products is a crucial for interpretability? Why does one product cant be just better than other two?

This sentence has been rephrased to « essential to maintain the comparability of the subsequent analyses based on these products »: indeed, consistency is not crucial for interpretability. But as the products are already used for studies and publications, it is important to know how we can compare these studies.

Line 660 – if the lower quality of data would be chosen – coverage factor would be greater, isn’t it like that? This both parameters are basically inverse dependent. Please give a measure that change of data quality filter affects the emission distribution. It wasn’t in the result, beside the counterintuitive statement that larger errors results in enhancements of the emissions.

The « high-quality » mention has been removed. In this study, the impact of quality filtering was quantified through the impact of coverage in the « Common » OSSEs. Based on our results, we can only provide a measure for these OSSEs (in terms of relative increment), but not on the real emission distribution.

Line 661 – what deeper understanding refers to? Please, provide the quantities for each factor in which it impacts the retrieved fluxes:

Quality filtering,

Spatial coverage,

Spatial distribution,

Uncertainty of observation,

Ability to constrain emission (not discussed earlier).

This has been discussed in the Sections 3.3 and 3.4, but cannot be summed up with a single quantification. In Section 3.4, the metrics is the relative increment and enables to quantify the impact of varying coverage and observation uncertainty on the resulting emission distribution, in terms of spatial and temporal constraint. The impact of differences in the distribution of XCH_4 has been discussed with the *Diff* OSSEs. It is rather a qualitative analysis than an absolute comparison of the impact of the mentioned factors on estimated emissions.

Line 665 – Application of ML to discover that retrievals are dependent on aerosol, albedo and striping is not new – all that was clearly reported in cited literature. Why it is not an object of this investigation? It is what I expected.

See previous answer to the Abstract specific comment. The ML application enabled to quantify the impact of these parameters on the differences of XCH_4 between products, which had not been done before (it had been done by Balasus et al., 2023, but in the case of the TROPOMI-GOSAT difference). The quantification of the impact of albedo, aerosols and striping on emissions is a very important perspective. We chose not to focus on these parameters in the OSSEs, as it would have required to modify the XCH_4 products (providing a correction). We wanted to use the products « as an user would do ». The perspective of the albedo/aerosols/striping impact on emissions has been highlighted as the most important future work.

Line 666 – Recent…. – why not to calculate how much those activities (developments) can improve the emissions coherence (this would be a great answer to the question).

This would be a very interesting study. However, it would require a detailed study comparing the previous versions of the products, that are not used anymore. It would not enable us to make predictions on future improvements. We prefered to give a snapshot comparison: a comparison of emissions estimated from inversions given the products available at the moment. Comparing the impact of the improvements goes beyond this study. The statement on the fact that future improvements should « improve the consistency of products » has been removed, as it is not supported by a result of this article.

Line 681 (probably most important one) – Without the uncertainties this numbers are not giving proper answer to title question. Please set the hypothesis and confirm or negate it with proper statistical analysis.

See the answer to general comments on the uncertainty considerations.

Author response v3 - 19/12/2025

Response to the Editor

We thank the editor for their comments and guidelines. We respond to them below: the comments are copied hereafter and shown in black, our author responses in blue. Changes in the manuscript are mentionned in green.

This work has accomplished a lot and produced both interesting and important findings. However, there are still some outstanding issues of concern which need to be addressed. Looking forward to your updated version, which works to address the following issues:

1. More depth with respect to model uncertainties and measurement uncertainties.
2. Comparisons with external observational datasets (which do exist)
3. A deeper explanation of the realism of the results compared with on the ground reality
4. Analysis probing/discussing what weaknesses currently may need to be addressed by future works.

A major issue arises from the fact that your observational data has both (a) far higher spatial and temporal resolution than your model, yet (b) is still coarse when compared with surface observations-especially from large point sources like coal mines.

Therefore a deeper analysis (or least discussion) needs to be done which considers how much information and/or precision are lost due to the aggregation of the observations into the model environment, subsequent modeling and its uncertainties, and the final disaggregation and comparison with on-the-ground reality. For example, if your model is relying on an emissions dataset which has "zero emissions" on a grid which in reality has a "large source", how is this uncertainty propagated through your system? I specifically raise this point because the reviewers have hinted that this is what may be occurring in actuality.

Overall, we have tried to address as much as possible the elements you mention, while keeping the focus of the study on the relative comparison of emissions obtained from the assimilation of satellite XCH4 datasets. We have added added references to relevant work and improved the discussions on several topics: the emission uncertainties and the ground reality, the need to bridge the gap between local studies and regional studies, the comparison with surface stations. We detail below these responses.

1. The question of uncertainties, for both model and measurements, is indeed a crucial question. In variational inversions, uncertainties are provided as inputs in the **B** and **R** matrix. The model error and the uncertainties on the a priori emissions is contained in the **B** matrix. The error is assumed to be 100% for CH4 emissions (see Section 2.4.2). Since the uncertainty is not provided in the a priori emission datasets, the errors have been chosen based on values from the intercomparison of CH4 inversions in Europe (Ioannidis et al., 2025), and from the study of Szénasi et al. (2021). The **R** matrix contains observation errors, and has been built with the individual retrieval errors provided in the satellite/surface observation datasets.

The uncertainties on posterior emissions are not directly accessible in our 4D-Var inversions. As discussed in the « Author response to referee #3 », this lack of uncertainty range intrinsically limits the statistical analysis of the consistency of posterior emissions. However, similar configurations of the CIF-CHIMERE inversion system have been used for all the inversions, thus we can address the differences of the spatial and temporal CH4 emission distributions due to the differences in TROPOMI observation products.

Also, the uncertainty reduction (independent of the observation vector) can be estimated through our OSSEs: the ratio of the standard deviation of priors and the one of posterior emissions across the ensemble samples is an estimation of 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. We do not use these values for setting error bars to our budget estimates, for two reasons: 1) the limited statistics of the ensemble (4 samples) and 2) the lack of proper uncertainty estimation for the prior (not provided with the emission inventories). We have added a discussion on this question in Section 3.4.1 and in the conclusion.

The propagation of the uncertainties in the model (such as your example of a large source not counted in the “zero emissions” pixel on the model grid), as well as the question of aggregation and disaggregation of the inputs/outputs of the model are crucial topics. However, addressing this complex issue would require a dedicated study, as the work presented in Zheng et al. (2025): we feel it goes beyond the scope of this manuscript. We study emission outputs at both pixel, country-scale and multi-country-scale (see Figure 10, Table S1): aggregating the results at larger scales smooth local errors that can arise from these problems. Our study insists on the fact that even at national scale, significant differences in CH4 emissions remain between the inversions based on the different TROPOMI products.

2./3. Comparisons with external datasets have been provided in several ways: comparison of the TROPOMI XCH4 observations with GOSAT observations (Section 2.1.3) and with TCCON observations (Annex C), comparison of the inversions and OSSEs of satellite observations with an inversion using independent surface observations (Sections 3.3, 3.4), evaluation of the CH4 simulated concentrations (prior and posterior) at these surface stations (Section 3.5).

Our study focuses on the relative comparison of TROPOMI satellite products, not on the absolute validation of emissions inferred from TROPOMI inversions. The latter has been analysed in the validation of global (Yu et al., 2023 ; East et al., 2025) or regional (Liang et al., 2023 ; Ioannidis et al., 2025) inversions with one TROPOMI product or the other, but few inter-comparisons of all the TROPOMI products had been performed : an issue that we address in this article.

Local studies provide valuable information on the emissions, and the comparison with regional/global studies is important. Integrating data from hotspots and high emission local areas in regional and global studies is a perspective for future work, and is made possible by satellites as TROPOMI. This is also why it is necessary to have a clear idea of the biases of the regional studies performed with satellite data, the issue that we tackle in this article.

For this study, we have made new comparisons in Poland and Romania as suggested by Referee #4 (see answer to question 5). However, for reasons mentioned in the response and as the manuscript is already quite long, we do not incluse these additional comparisons to surface datasets in the main text. The deeper analyses of the comparisons is mentionned in the text as future work.

4. In the conclusion of the manuscript, as well as in the ends of “Results” sub-sections, we discuss the limits of this analysis and future work :

- Regarding TROPOMI observations: we used the XCH4 datasets with as low post-processing as possible. Further understanding of the biases in the derived CH4 emissions requires corrections of the albedo and aerosols biases highlighted in Section 3.1.
- The characterization of uncertainties related to the background (lateral and top boundary conditions, stratosphere) and its role in the optimization process (Section 3.3). This has been the object of a recent study (Nesser et al., 2025).

- The gap between satellite-based and surface-based inversions, in terms of space-time distribution of the increments and of simulated CH4 measurements (Sections 3.3, 3.5).

5. Regarding the aggregation of observational data with both a higher space-time resolution that the model and a coarser resolution than surface data, two elements can be mentionned.

First, we have performed inversions with super-observations, aggregating observations at the spatial resolution of the model (0.5°x0.5°) and with hourly resolution. The posterior emissions and the simulated observations (prior and posterior) were close (<1% difference), while the patterns of the spatial and temporal distributions remain unchanged. For this reason, we chose to use the entire dataset to remain as general as possible, as “a TROPOMI user would do”.

Secondly, the coarser resolution than surface data limits the analyses at local scale: this is why we perform analyses at both the pixel and the country-scale. The objective is not to focus on local point sources, but to include the emissions at larger scales to build country-scale and regional-scale budgets, a complementary approach to local studies.

References:

East, J.D., Jacob, D.J., Jervis, D. et al. Worldwide inference of national methane emissions by inversion of satellite observations with UNFCCC prior estimates. Nat Commun 16, 11004 (2025). https://doi.org/10.1038/s41467-025-67122-8

Ioannidis, E., Meesters, A., Steiner, M., Brunner, D., Reum, F., Pison, I., Berchet, A., Thompson, R., Sollum, E., Koch, F.-T., Gerbig, C., Wang, F., Maksyutov, S., Tsuruta, A., Tenkanen, M., Aalto, T., Monteil, G., Lin, H., Ren, G., Scholze, M., and Houweling, S.: An inter-comparison of inverse models for estimating European CH4 emissions, Earth Syst. Sci. Data Discuss. [preprint], https://doi.org/10.5194/essd-2025-235, in review, 2025.

Liang, R., Zhang, Y., Chen, W., Zhang, P., Liu, J., Chen, C., Mao, H., Shen, G., Qu, Z., Chen, Z., Zhou, M., Wang, P., Parker, R. J., Boesch, H., Lorente, A., Maasakkers, J. D., and Aben, I.: East Asian methane emissions inferred from high-resolution inversions of GOSAT and TROPOMI observations: a comparative and evaluative analysis, Atmos. Chem. Phys., 23, 8039–8057, https://doi.org/10.5194/acp-23-8039-2023, 2023

Nesser, H., Bowman, K. W., Thill, M. D., Varon, D. J., Randles, C. A., Tewari, A., Cardoso-Saldaña, F. J., Reidy, E., Maasakkers, J. D., and Jacob, D. J.: Predicting and correcting the influence of boundary conditions in regional inverse analyses, Geosci. Model Dev., 18, 9279–9291, https://doi.org/10.5194/gmd-18-9279-2025, 2025.

Yu, X., Millet, D. B., Henze, D. K., Turner, A. J., Delgado, A. L., Bloom, A. A., and Sheng, J.: A high-resolution satellite-based map of global methane emissions reveals missing wetland, fossil fuel, and monsoon sources, Atmos. Chem. Phys., 23, 3325–3346, https://doi.org/10.5194/acp-23-3325-2023, 2023

Zheng, B., Cohen, J. B., Lu, L., Hu, W., Tiwari, P., Lolli, S., Garzelli, A., Su, H., and Qin, K.: How can we trust TROPOMI based Methane Emissions Estimation: Calculating Emissions over Unidentified Source Regions, EGUsphere [preprint], https://doi.org/10.5194/egusphere-2025-1446, 2025.

Response to RC4

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.

This study compares the emissions estimated from variational inversions in 2019 over Europe with a resolution of 0.5˚ assimilating three TROPOMI products of dry-column methane mole fractions (XCH4), which is a valuable and an important work. The results shows that the largest contributions to XCH4 differences may be attributed to aerosol scattering and sensitivity to albedo. And find that seasonal emissions are highly correlated across the inversions. Besides, this study also used the Observing System Simulation Experiments (OSSEs) to disentangle the drivers of differences between the posterior emissions, and reveal that observation density and errors, averaging kernels and prior profiles play a key role in the inversion's capacity to constrain the emissions. Overall, the insights of this study are valuable, and the conclusions drawn are helpful to other studies. I recommend publication in ACP after addressing the following general and specific comments.

General comments

1. Line 178, "average concentrations decrease slightly from January to May, reaching a minimum in April/May", does this pattern have anything to do with the monthly observations? Since from figure 1(a), the observation count decreased from April, and reaching a minimum in May. From Figure 2(a), monthly XCH4 reach a peak in November, while from figure 1(a), there is a minimum observation count in November, how does the observation (sampling difference) affect the results?

The sampling difference indeed affects the temporal and spatial averages of XCH4. The peak in November is partially explained by the decrease in observation count for this month, and especially in the low observation count in the northern latitudes where XCH4 is lower.

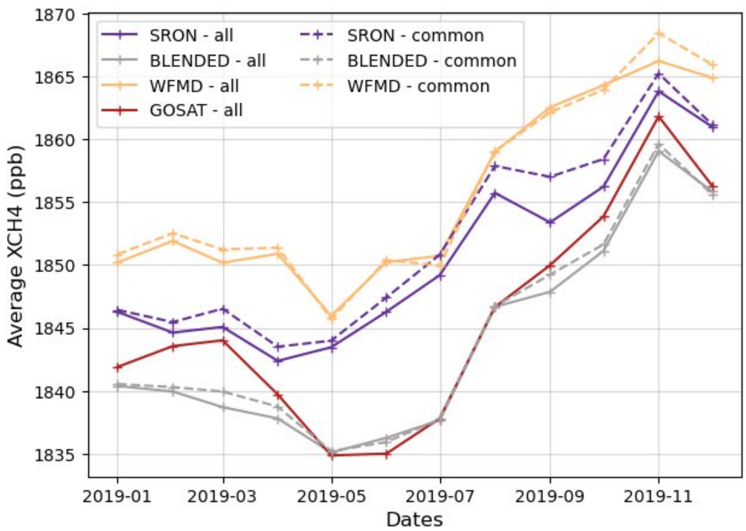

Figure – Similar to Figure 2 (a), including the TROPOMI products restricted to common observations (dashed lines).

To disentangle the effect of the differences in XCH4 variations and in the sampling variations, we compare the TROPOMI products restricted to the common observations across the 3 products (see the remark L.176 on the average values). The time variations are compared in the figure below, they slightly differ between « all observations » and « common observations » but overall the shapes remain similar. A remark has been added in the manuscript to clarify this point.

2. As can be seen from Figure 10(c), only the satellite has a seasonal change, but there is no obvious change between the ground observation and the priori, and the change trend of the surface observation from March to June is not the same as the change trend of the satellite. How can the accuracy of the satellite results be verified if both surface observations and prior comparisons are inconsistent with the satellite results? How to verify the effectiveness of the product?

In this study, the inversion assimilating independent surface observations (from the ICOS, WDCGG, NOAA ESRL and EBAS databases, see Section 2.1.2) is used for comparison. The TROPOMI-based and surface-based emissions show higher emissions in winter and a minimum in summer: satellite and surface inversions are rather consistent, but they are not consistent with the prior. However, the satellite-based inversions show a peak in April and May that is not seen neither in the surface-based posterior emissions nor in the prior emissions. As mentionned L.465, the origin of this peak has been investigated but we provide only partial explanations.

The main focus of this article is on relative differences between products, thus the comparison of the results of surface and satellite inversions has neot been extensively analysed. It has been pinpointed as a challenge for future work, to reconcile the "satellite perspective" and the "surface perspective". Indeed, surface and satellite observations have intrinsic differences (in-situ vs. remote sensing, localized vs. global, uncertainties...). These differences lead to different constraints on emissions in the inversions. The consistency of satellite-based and surface-based inversions is necessary to evaluate the realism of posterior CH_4 emissions.

This challenge is also highlighted by the comparison of prior/posterior observations at the surface stations in Section 3.5 (see comment #6). The analysis should be extended to a larger set of prior emission and observation datasets (other satellite/surface datasets). The use of ground-based remote sensing observations (such as TCCON or COCCON) could also provide valuable data to bridge the gap between surface and satellite inversions. We have improved the discussion in the manuscript, but an extensive dedicated study would go beyond the scope of the article.

3. Figure S3 reveals substantial spatial variability which on a pixel-by-pixel basis is a lot larger than the changes on a country-by-country basis given in Table S1. This needs to be explained more deeply. Some of these changes look to be 100% or more. Why has the optimization not captured these very well? What impacts could this have for new or emerging emissions sources? Is the optimization weighted individually pixel-by-pixel? Could there be some issues in the a priori spatial temporal distribution, and therefore the model error is much larger than currently considered?

Figure S3 shows the absolute increments for the OSSE scenarios, while Table S1 shows increments on total country emissions for the 4 inversions. These two figures should not be compared : synthetic pseudo-observations are assimilated in the OSSEs, while real observations are assimilated in the inversions. Moreover, increments are shown as absolute values in Figure S3 to capture their intensity. But country emission budgets in the inversions are the sum of positive and negative increments, thus smoothing high pixel-to-pixel variations. So I'm not sure I understand the objective of your first question.

The optimization is weighted individually pixel-by-pixel, with spatial and temporal correlations of errors (see Section 2.4.2). The model error is set to 100% for CH_4 emissions. Uncertainties are not provided in the a priori inventories, so the errors have been chosen based on values from the intercomparison of CH4 inversions in Europe (Ioannidis et al., 2025), and Szénasi et al. (2021).

References:

Ioannidis, E., Meesters, A., Steiner, M., Brunner, D., Reum, F., Pison, I., Berchet, A., Thompson, R., Sollum, E., Koch, F.-T., Gerbig, C., Wang, F., Maksyutov, S., Tsuruta, A., Tenkanen, M., Aalto, T., Monteil, G., Lin, H., Ren, G., Scholze, M., and Houweling, S.: An inter-comparison of inverse models for estimating European CH4 emissions, Earth Syst. Sci. Data Discuss. [preprint], https://doi.org/10.5194/essd-2025-235, in review, 2025.

Szénási, B., Berchet, A., Broquet, G., Segers, A., van der Gon, H. D., Krol, M., Hullegie, J. J., Kiesow, A., Günther, D., Petrescu, A. M. R., Saunois, M., Bousquet, P., and Pison, I.: A pragmatic protocol for characterising errors in atmospheric inversions of methane emissions over Europe, Tellus B: Chemical and Physical Meteorology, 73, 1–23, https://doi.org/10.1080/16000889.2021.1914989, 2021.

4. Figure S1 reveals an observation count. However, I am still not sure about how many days' worth of observations there are in each grid. For example, if a model-sized grid has one TROPOMI pixel each day for 10 days, and another model-sized grid has ten TROPOMI pixels on a single day, then would both cases count as having the same observation count? This map is confusing because it is merging together the spatial and temporal counts. In reality, the value of these two different cases would be very extensive. They would have substantial differences in terms of wind, topography, aerosols, clouds, surface reflectance, and other issues. I am not sure if I can believe the model uncertainty until this can be disentangled.

This map has been though to give the reader an idea of the variations of the spatial distribution of the observation count along the year, and how they compare between products. Indeed, in both cases you mention, the 10 observations would be counted the same way : space and time counts cannot be fully disentangled in one map. A second set of maps (Figure S1b) has been added, showing the number of days with at least one observation per model pixel. The combination of both figures gives a broader understanding of the observation count distribution.

However, the content of the map is not what is provided as an input to the model: the model takes into account each observation individually, with the associated date and location, thus the associated topography, wind, etc. This figure is not related to the model uncertainty.

5. I noticed that you do not use any of the surface observation constrained emissions calculated from coal mines in Poland (Tu et al., 2022). How do your computed emissions compare with these results? Since they are independent from your inversion, such a comparison would provide a truly independent comparison.

Comparing the emissions from our inversions to the ones from coal mines in Poland would provide a relevant comparison (not independent though, as TROPOMI XCH4 is used to constrain the emissions in Tu et al., 2022). We propose the two following comparisons at local scale, in Poland and Romania. Please note that the coal mining (Poland) and oil and gas production (Romania) emissions can vary from one year to another, making it hard to compare different years.

Poland: in the Upper Silesian Coal Basin in Poland (49.3-50.8°N, 18-20°E), emissions retrieved from TROPOMI XCH4 (SRON product) and TROPOMI + IASI TXCH4 are respectively about 496 kt/yr and 437 kt/yr, and they are compared to the coal mining emissions from the E-PRTR inventory (448 kt/yr) and the CoMet inventory (555 kt/yr), for the year 2021.

For the same region (see figure below), total methane missions in 2019 are 722 kt/yr for the prior (coal mining emissions are from EDGAR v8), 762 kt/yr for the posterior of SRON, 670 kt/yr for BLENDED, 289 kt/yr for WFMD and 638 kt/yr for the surface-based inversion.
Except for WFMD, total emissions are higher than those from Tu et al. (2022), because the prior of EDGAR has higher emissions. While SRON leads to increased posterior emissions, the Surface and BLENDED inversions are consistent and lead to total emissions closer to the results of Tu et al. (2022). WFMD leads to very low emissions in comparison to the others, as the specific pixels of this region have very strong negative increments. These locally high gradients may be unrealistic. This is also the case for the next case in Romania (see the increments on Figure 10a).

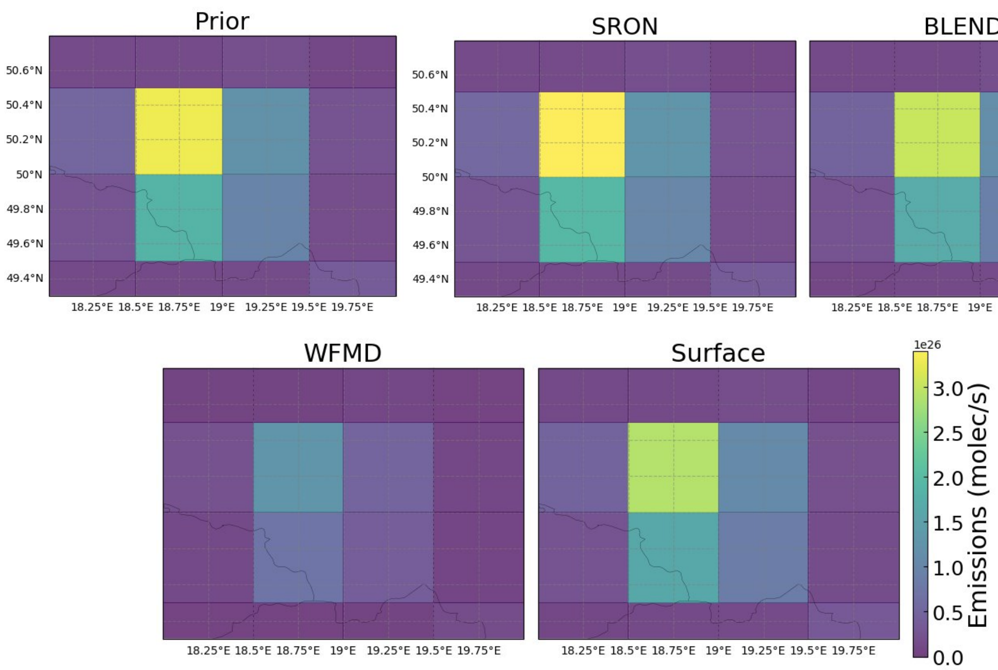

Figure – Maps of prior and posterior CH4 emissions for the 4 inversions performed in this study, over the USCB region studied in Tu et al. (2022), in 2019. Emissions are in molec/s.

<u>Romania:</u> in the study area of the ROMEO campaign in Romania, emissions retrieved from the airborne measurements are respectively about 227 kt/yr in 2019 (Maazallahi et al., 2025) and from 76 to 175 kt/yr in 2021 (four scenarios in Kuhlmann et al., 2025), to compare to 71kt/yr in 2019 and 66 kt/yr (national inventory report for UNFCCC), and 80 kt/yr for IEA.
For a similar region (44-45.5°N, 23-27°E), total methane missions in 2019 are 292 kt/yr for the prior (mainly from EDGAR v8), 256 kt/yr for the posterior of SRON, 240 kt/yr for BLENDED, 56 kt/yr for WFMD and 235 kt/yr for the surface-based inversion.
Except for WFMD, every inversions make emissions closer to the values of Kuhlmann et al. and Maazallahi et al., and once again BLENDED and Surface inversions are consistent in this area.

These comparisons are limited by the following arguments:

- Extrapolating the comparison to different years may introduce errors.
- The area is not clearly defined for Romania, and the studied areas contain only a few 0.5°x0.5° pixel, making local estimations very sensitive to high pixel increments. This is why we suggest to focus on comparisons at the country scale, rather than the local scale.
- Our CH4 emissions contain all the anthropogenic and natural emissions. Even if the emissions are dominated by coal mining in Poland, and oil/gas production in Romania, comparing totals is theoretically not correct.

For these reasons, and also as we already have included comparison with surface measurements, we have chosen not to add these comparisons in the manuscript. The manuscript is already long, and we want to keep the focus on the relative differences between emissions from the different TROPOMI products, to avoid making the paper unclear. We mention the following references for future work: a deeper and dedicated analysis is required to make a proper and relevant comparison and to validate the emission outputs of satellite-based inversions against surface measurements.

References :

Kuhlmann, G., Stavropoulou, F., Schwietzke, S., Zavala-Araiza, D., Thorpe, A., Hueni, A., Emmenegger, L., Calcan, A., Röckmann, T., and Brunner, D.: Evidence of successful methane mitigation in one of Europe's most important oil production region, Atmos. Chem. Phys., 25, 5371–5385, https://doi.org/10.5194/acp-25-5371-2025, 2025.

Maazallahi, H., Stavropoulou, F., Sutanto, S. J., Steiner, M., Brunner, D., Mertens, M., Jöckel, P., Visschedijk, A., Denier van der Gon, H., Dellaert, S., Velandia Salinas, N., Schwietzke, S., Zavala-Araiza, D., Ghemulet, S., Pana, A., Ardelean, M., Corbu, M., Calcan, A., Conley, S. A., Smith, M. L., and Röckmann, T.: Airborne in situ quantification of methane emissions from oil and gas production in Romania, Atmos. Chem. Phys., 25, 1497–1511, https://doi.org/10.5194/acp-25-1497-2025, 2025.

Tu, Q., Schneider, M., Hase, F., Khosrawi, F., Ertl, B., Necki, J., Dubravica, D., Diekmann, C. J., Blumenstock, T., and Fang, D.: Quantifying CH4 emissions in hard coal mines from TROPOMI and IASI observations using the wind-assigned anomaly method, Atmos. Chem. Phys., 22, 9747–9765, https://doi.org/10.5194/acp-22-9747-2022, 2022.

6. My largest worry/concern is that the inverse optimized emissions compare worse against surface observations than the non-optimized emissions. This point needs to be explored more deeply.Lines 608-611 and associated tables/figures/etc.

As presented in Table 6 and Figure 13, the concentrations simulated with the posterior emissions are not clearly improved by the inversion: they are slightly improved for BLENDED (lower bias and RMSE, fit between observed and simulated concentrations improved for 53% of the stations), but degraded for WFMD and SRON. We have pinpointed the gap between the satellite and surface “points of view”: optimizing emissions with respect to satellite observations does not imply a better fit for surface concentrations. This is an important issue for the comparability of satellite-based and surface-based studies. Comments have been added in the text and in the conclusion. A deeper analysis is indeed required to improve the consistency of these categories of inversions, but we feel it goes beyond the scope of our work, which focuses on satellite XCH4 datasets.

7. Figures 11a and 10c also point out issues which occur when analyzing emissions at higher temporal frequency. I would like to see how these vary at weekly or day-to-day comparison.

The two figures below are similar to 10c and 11a, but at a weekly frequency. The control vector is optimized at a weekly resolution (in both the inversions and OSSEs) and a priori emissions have a monthly resolution, thus the day-to-day variations do not provide additional information.

Fig 10c: the weekly and monthly curves have similar shapes, with more noise especially for the *WFMD* and *Surface* inversions. The discussion of the comparison of temporal variations between the four inversions remains similar.

Fig 11a: the weekly variations show a lot more noise than the monthly ones. We are not so interested in the absolute variations, but rather in the relative differences with one scenario or another. From this point of view, the relative « positions » of the curves are overall similar and lead to the same analyses as presented in the manuscript. A sentence has been added in Section

3.4.1 to mention the fact that monthly variations smooth higher frequency variations, but relative differences are overall conserved. For the sake of readability, we prefer to keep the monthly variations in the manuscript.

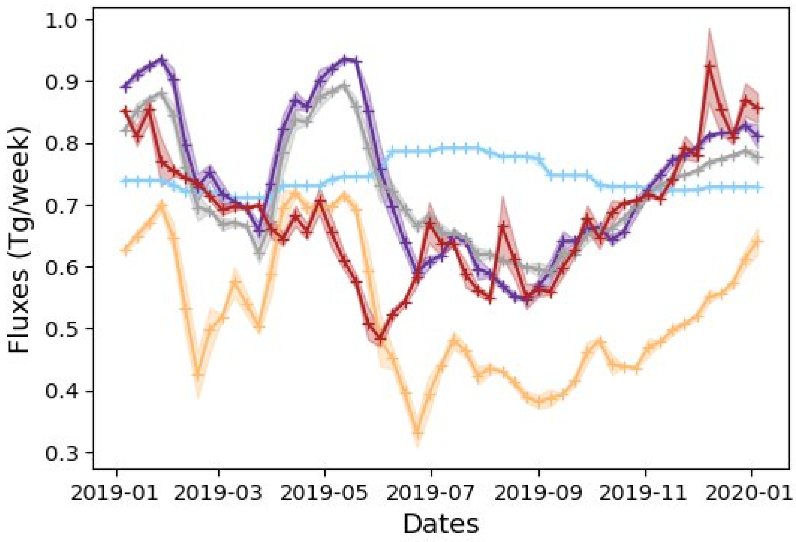

Figure – Similar to Figure 10 (c), but at a weekly temporal resolution.

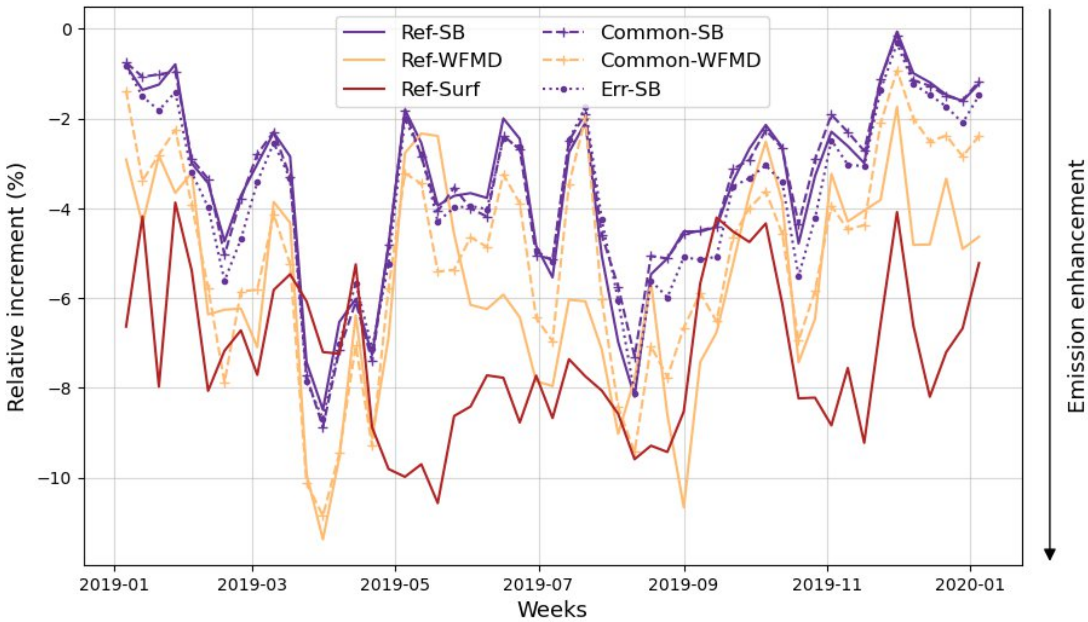

Figure – Similar to Figure 11 (a), but at a weekly temporal resolution.

8. I am concerned that your CH4 retrieval uncertainty is too low, based on your own statement on lines 143-146 and on lines 199-200. There have been multiple recent papers pointing out that there is a substantial impact of aerosols on the TOA and upward radiances in this region based on studies using AERONET (Tiwari et al., 2023), MISR (Liu et al., 2024), and TROPOMI and AEROENT (Tiwari et al., 2025). Furthermore, there are studies which also demonstrate that there are substantial impacts of aerosols on the ATM and hence absorption radiances in this region based on using sunphotometer and aerosol microphysical observations in tandem (Guan et al., 2025; Yang et al., 2025). A larger uncertainty on the column retrieval will lead to a change in the emissions by changing the spatial and temporal variation of emissions (Lu et

al., 2025), as well as call into question both very high and very low emissions values (Zheng et al., 2025).

The purpose of our study is to compare emissions from inversions assimilating TROPOMI datasets as close as « a data user would do ». Therefore, we do not try to improve the retrieval uncertainty: they are only modified in the OSSE « SB-Err », to evaluate the impact of the inconsistencies between errors of SRON/BLENDED and WFMD.
Also, increasing the uncertainty would only make the posterior closer to the prior. It would not address the biases of the observation datasets and their impact on the inversion results, which we explore in this study. For these reasons, we do not modify the XCH4 uncertainties.

However, the papers you mention are very important in terms of the sources of differences between products and the analysis of optimized posterior emissions. The substantial impact of aerosols is consistent with the analysis of Section 3.1. A discussion on the retrieval uncertainties and the inferred emissions in Section 2.1.3.

9. Your simple comment about the change in the OH dominating the changes in the CH4 on lines 181-182 requires a deeper set of explanations. New work demonstrates that increases in coal mining (and hence CH4 emissions) are paired with increases in CO and BC emissions (from steel and iron production, as well as biomass burning)(Frausto-Vicencio et al., 2023; Mallet et al., 2024), and hence these other co-emitted species will also impact the lifetime of CH4 (by competing with OH radicals) during this period of time (Hu et al., 2024; Li et al., 2025; Wang et al., 2025; Deng et al., 2024; Pitt et al., 2019).

This sentence had been written to explain the XCH4 time variations of Fig. 2a, which are related to the hemispheric seasonal cycle of CH4 in the atmosphere. We agree that the work you mention shows that the competition between chemical species is more complex than presented here.

Integrating a discussion would go out of scope of the study, because chemistry is neglected in the transport model due to the small ventilation time scale of the domain (~10 days). Therefore, we remove this oversimplified comment from the text.

Author response v4 - 24/04/2026

Response to the Editor

We thank the editor for their comments and feedbacks. We respond to them below: the comments are copied hereafter and shown in black, our author responses in blue.

You have done substantial work responding to the previous rounds of review. Two reviewers are satisfied. The other two reviewers acknowledge your efforts as well, yet raise important questions that need to be addressed before publication.

A tension revealed by your study, that inversions optimized for satellite observations do not always improve agreement with surface observations, is important. It is also, as you have noted note, not fully resolved. This is not a flaw; it is honest, and the community needs to understand. The reviewers have asked you to go further in exploring why this is the case and what it means for the usability of your products.

Please address the following points in your revision:

Model errors: The reviewers ask for greater clarity about how model errors (transport, boundary layer representation, tropopause height) may interact with retrieval biases. You discuss these separately, but the question is how they work in tandem. A more explicit discussion of this coupling, even if it is speculative, would help readers understand what inversions can deliver and what their limitations are.

A clarification on the model errors has been added as a response to RC3 and in the main text. It is a broad and complex topic, with a limited number of quantitative analyses. There are multiple links between retrieval errors and the atmospheric conditions as well as between the atmospheric transport model errors and the atmospheric conditions: correlation between retrieval and transport errors for a given pixel at a specific date; similarity of horizontal spatial patterns and temporal patterns of errors; impact of vertical transport on errors (particularly important for CH_4 is the representation of stratosphere-troposphere exchanges and the tropopause height). The high complexity and multiplicity of these relationships makes this topic extremely difficult to tackle, and even to discuss in a relevant way.

As an example, a specific air mass with relatively high aerosol and methane concentrations can generate high retrieval errors over its horizontal extent while the transport model could erroneously shifts this extent: in this case, it would be extremely difficult to build a relevant discussion on the coupling between the retrieval and atmospheric transport errors. How closely do the spatial patterns of transport model and retrieval errors follow the horizontal patterns of the atmospheric dynamics, and with which level of correlations between the two types of errors ?

The derivation of suitable statistics of the atmospheric transport model errors is already a highly complex topic. In practice, it is based on ensembles of simulations with varying models and input parameters (Locatelli et al., 2013, Szénási et al., 2021), or using the residual error method (Heald et al., 2004, Sheng et al., 2018, Chen et al., 2023). The results are generally summarized as statistics and the analyses hardly attempt at mapping the errors or to characterize them as a function of the dynamical patterns, so that we can hardly confront maps of transport errors with maps of retrieval errors for a practical inference of their correlation.

The derivation of spatial correlations in the model errors is therefore highly challenging, and these errors are usually (like in this study) modeled without correlations in the inversion configurations.

Here, **R** is diagonal and without model errors. Even if model errors were considered, the **R** components would be the sum of an observation error $\mathbf{R}_{obs}$ and a model error $\mathbf{R}_{mod}$, without correlation terms. These errors would be independent, thus adding an average model error would not change patterns in the outputs of the inversions, only reduce the fit to the observations. The analysis of the potential spatial coherence between the model errors and the retrieval errors would thus hardly be exploited in the current inversion configurations. This complex topic goes beyond the scope of our study.

Spatial and temporal validation at native scales: One reviewer's request for a grid-by-grid, time-by-time comparison is well-founded. You argue aggregation smooths local errors, but such smoothly also obscures the reliability of the retrievals at the scale needed for attribution. A representative analysis (for example, scatterplots for a few key regions or seasons) would help the community understand the magnitude of local mismatches and whether the spatial patterns are reasonable.

See responses to RC4.

Attribution: The goal of this work is to support mitigation efforts. If the spatial distribution is uncertain, even though the aggregates look good, what does this imply about using these inverse products for attribution of specific emissions sources or regions? A frank discussion, even just a rough quantification of the limits of confidence in attribution, would strengthen the paper's contribution.

See responses to RC4. A discussion to the previous point and this one have been added in the manuscript. Additional figures have also been added.

Seasonality: One reviewer asks whether the spring peak in satellite inversions appears in TCCON comparisons, and this is worth pursuing. Even a brief seasonal analysis for a few key sites (as suggested) would clarify whether this is real or an artifact.

The analysis has been performed for a few TCCON sites (see the response to RC3). The associated figures have been added in Supplementary material, and a discussion on the issue has been added in the text.

I recognize that some of these questions may not have complete answers. You are working at the frontier of what is possible when interacting with such a complex system. The goal is not to resolve everything, but to be transparent about what we know, what we don't know, and what this means for end users of your products.

Thank you again for your careful and thoughtful work. I look forward to seeing the revised manuscript.

References:

Chen, Z., Jacob, D. J., Gautam, R., Omara, M., Stavins, R. N., Stowe, R. C., Nesser, H., Sulprizio, M. P., Lorente, A., Varon, D. J., Lu, X., Shen, L., Qu, Z., Pendergrass, D. C., and Hancock, S.: Satellite quantification of methane emissions and oil–gas methane intensities from individual countries in the Middle East and North Africa: implications for climate action, Atmos. Chem. Phys., 23, 5945–5967, https://doi.org/10.5194/acp-23-5945-2023, 2023.

Heald, C. L., D. J.Jacob, D. B. A.Jones, P. I.Palmer, J. A.Logan, D. G.Streets, G. W.Sachse, J. C.Gille, R. N.Hoffman, and T.Nehrkorn (2004), Comparative inverse analysis of satellite

(MOPITT) and aircraft (TRACE-P) observations to estimate Asian sources of carbon monoxide, J. Geophys. Res., 109, D23306, doi:10.1029/2004JD005185.

Locatelli, R., Bousquet, P., Chevallier, F., Fortems-Cheney, A., Szopa, S., Saunois, M., Agusti-Panareda, A., Bergmann, D., Bian, H., Cameron-Smith, P., Chipperfield, M. P., Gloor, E., Houweling, S., Kawa, S. R., Krol, M., Patra, P. K., Prinn, R. G., Rigby, M., Saito, R., and Wilson, C.: Impact of transport model errors on the global and regional methane emissions estimated by inverse modelling, Atmos. Chem. Phys., 13, 9917–9937, https://doi.org/10.5194/acp-13-9917-2013, 2013.

Sheng, J.-X., Jacob, D. J., Maasakkers, J. D., Zhang, Y., and Sulprizio, M. P.: Comparative analysis of low-Earth orbit (TROPOMI) and geostationary (GeoCARB, GEO-CAPE) satellite instruments for constraining methane emissions on fine regional scales: application to the Southeast US, Atmos. Meas. Tech., 11, 6379–6388, https://doi.org/10.5194/amt-11-6379-2018, 2018.

Szénási, B., Berchet, A., Broquet, G., Segers, A., Denier van der Gon, H., Krol, M., … Pison, I. (2021). A pragmatic protocol for characterising errors in atmospheric inversions of methane emissions over Europe. Tellus B: Chemical and Physical Meteorology, 73(1), 1–23. https://doi.org/10.1080/16000889.2021.1914989

Response to RC3

We thank the reviewer for their comments and feedbacks on the manuscript. We respond to them below: the comments are copied hereafter and shown in black, our author responses in blue.

General comments:

The study looks at the use of 3 different Tropomi XCH4 retrieval products in regional methane flux inversion for Europe. Authors embarked on solving a set of complex problems involving the biases of the retrieval products, transport model errors and uncertainties of the inverse model results. They try to trace down the flux differences to the drivers of XCH4 biases in the respective products, ending up in a complex analysis, involving use of ML approach following Balasus et al. 2023. As the reviewers of the original manuscript mentioned, the resulting report was not easy to read and extract the essential results. The problem was not fully resolved in the revised manuscript, but as noted by review reports, the study does contain a lot of useful analysis and is very much worthy of publication.

Detailed comments:

Authors faced number of complications to the analysis origination in several problems among which only some could be addressed.For clarity and readers' convenience it would be useful to state what is investigated in detail (eg contribution of aerosols and other variables to L2 biases), and what is not, such as transport model errors and biases. Among those, one can count on contribution of the tropopause height variation impacting observed XCH4 (eg Saad et al 2014) that could be miscalculated by CHIMERE.

References: Saad, K. M., et al: Derivation of tropospheric methane from TCCON CH4 and HF total column observations, Atmos. Meas. Tech., 7, 2907–2918, https://doi.org/10.5194/amt-7-2907-2014, 2014.

Following your recommendation, a list of the problems that are addressed in the study has been added. It is first mentioned in the introduction of the manuscript, and further detailed in Section 3.3. It is finally discussed in the conclusion.

The main objective of this study is to assess the consistency of the emissions estimated from inversions using 3 TROPOMI products, and how the differences in the products impact the spatial and temporal patterns of the emission estimates. We provide the following analyses:

- We compare the results of the inversions using the 3 TROPOMI products. We investigate the differences in the posterior emissions by describing the differences between products (coverage, XCH4 differences, observation errors, column averaging kernel, prior CH4 profile), and evaluating the contributions of aerosols, albedo, striping patterns and other retrieval variables to the XCH4 differences.
- Using OSSEs, we try to disentangle these drivers and quantify the impact of the differences in coverage, observation errors and XCH4 differences on the emissions. We also evaluate the impact of the background optimization on the posterior emission results, as part of the inversion system.
- We compare the amplitude of the differences with external evaluation (surface station measurements, TCCON observations).

Some limits of the analyses are highlighted as potential causes of discrepancy, but not deeply investigated:

- Albedo and aerosol-related biases on the estimated emissions are discussed qualitatively (paragraph « SRON-WFMD XCH4 differences » of section 3.4.2 partially addresses this issue as SRON and WFMD treat differently such biases) but not quantitatively estimated. It is suggested as a next step.
- A detailed study of the transport model errors (miscalculation of the tropopause height, representation of the boundary layer, transport errors at meso-scale...) and on how they interact with the errors related to the retrievals is not performed, as it is a complex topic that goes beyond the scope of the study (see response to the editor).

Fig. 4 Caption. Here one can use “Column averaging kernel” instead of simply “Averaging kernel”

The change has been done, in the figure caption and in the text.

Fig 10c Inverse modeling points to difference between Tropomi and surface data inversion in spring, could it be also found in TCCON comparison? Suggest adding seasonally resolving version of Fig C (for some key sites could be be OK) The domain average XCH4 seasonality on Fig 2a, does show same difference between products, which is not reflected in fluxes, as products with different seasonality happen to produce similar fluxes.

The peak of high CH4 fluxes seen in Fig. 10c in the results of the 3 inversions using TROPOMI data cannot be seen in the corresponding XCH4 observations (Fig. 2a), as mentioned in Section 3.3. Investigations were carried out to identify the cause of the positive corrections in Western and Central Europe associated with this peak, regarding the observations and the background optimization. Yet, it has not been fully clarified.

To compare the spring variations of XCH4 observations and TCCON total column measurements, a seasonal version of Fig. C1 has been added in the supplementary material (Figures S5, S6). A short discussion has been added in the corresponding Annex C, where the comparison TROPOMI/TCCON is made. The figure is shown below (Fig. 1) and includes the sites of Garmisch, Karlsruhe, Orléans, Paris and Sodankyla, where they are a sufficient number of co-located observations (except in JFM for Sodankyla).

Figure 1 indicates a low seasonal dependency in the differences, apart from Sodankylä. In spring (AMJ), the TROPOMI-TCCON difference is in average always positive (except for BLENDED in Paris, slightly negative), which indicates that TROPOMI XCH4 estimates are higher than the ones of the TCCON stations. However, this is also the case in winter (JFM), whereas the fluxes from inversions are not peaking in February and March 2019. A more detailed look at monthly resolution in Figure 2 shows high positive XCH4 difference in April in Karlsruhe and Sodankylä, but no clear evidence of a peak of the TROPOMI-TCCON difference in the April-May period. In Sodankylä, the TROPOMI-TCCON difference is positive during spring 2019, and negative in autumn 2019, consistent with the findings of Lindqvist et al. (2024).

Therefore, the comparison between TROPOMI and TCCON suggests positive difference in winter and spring, thus one would expect higher fluxes from TROPOMI-based inversions in comparison to a TCCON-based inversion. But the sping peak cannot be clearly identified in this comparison, suggesting the peak’s origin is partly related to the inversion system in itself.

<u>References:</u>

Lindqvist H, et al. Evaluation of Sentinel-5P TROPOMI Methane Observations at Northern High Latitudes. Remote Sensing. 2024; 16(16):2979. https://doi.org/10.3390/rs16162979

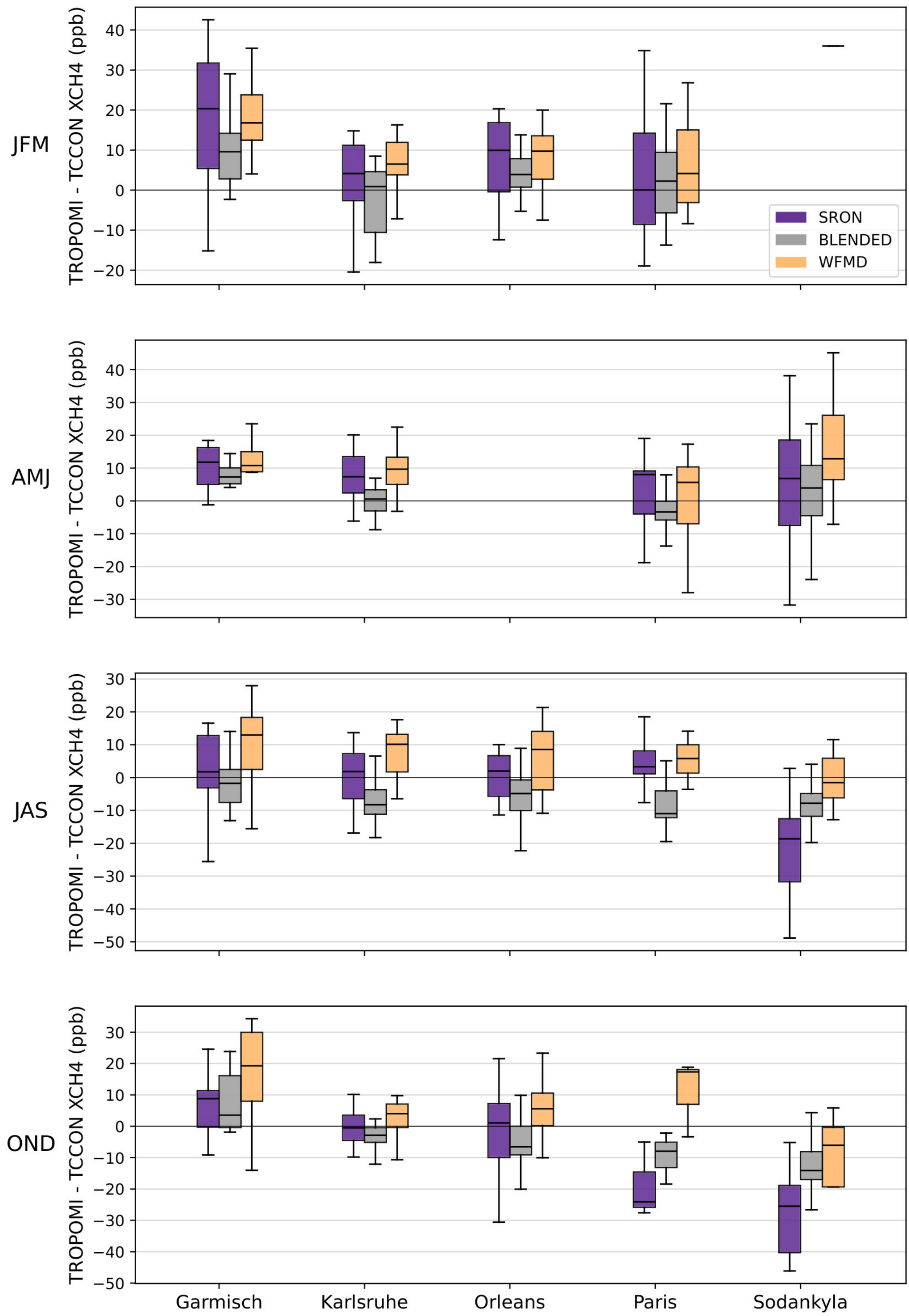

Figure 1: Similar as Fig. C1, but seasonally-resolved. Median and quartiles of the daily averaged differences between TROPOMI and TCCON XCH4 (ppb) for 5 TCCON stations, for all seasons in 2019.

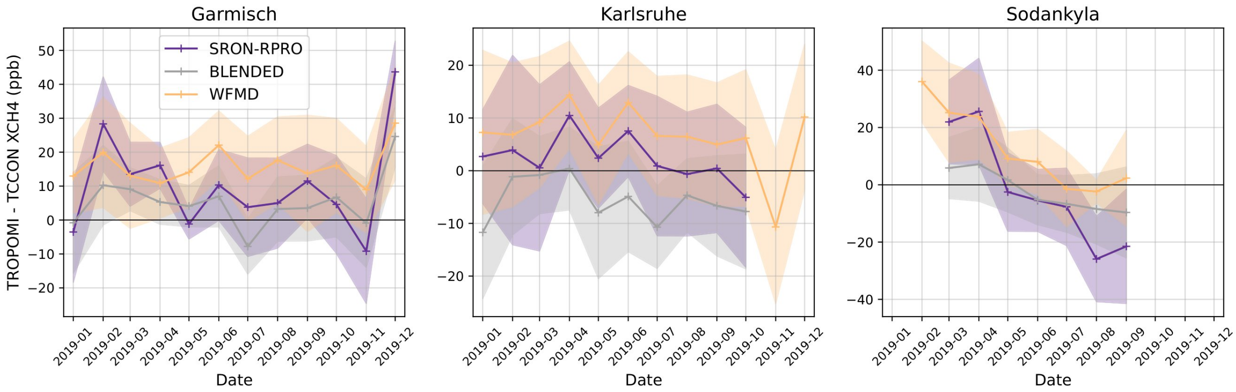

Figure 2: Time series of the monthly averaged differences between TROPOMI and TCCON XCH4 (ppb), for 3 stations. The shaded area shows the standard deviation within each month.

Response to RC4

We thank the author for their comments and for the discussion. We respond to them below: the comments are copied hereafter and shown in black, our author responses in blue.

This study provides a valuable reference for regional emission quantification driven by satellite data by comparing the inversion results of three TROPOMI products. The existing work is relatively complete, and the previously asked questions have been addressed reasonably well. However, there are some outstanding points which must be considered further.

1. The only way to get a national scale or regional scale budget is to sum all of the independent sources together. In your case, your model has a resolution of 0.5 x 0.5.You need to compare day-by-day, 0.5 x 0.5 grids with existing known emissions from medium and large surface sources.

We disagree with this statement. If the only way to get a national scale or regional scale budget were to sum all of the independent sources together, all CH4 atmospheric inverse modelling studies and efforts to develop new CH4 satellite missions should focus on point source studies with very high resolution images of the plume downwind to these sources. The local analysis is definitely the most sensible for large point sources with well identified plumes in the satellite images. However, the inverse modelling community has also shown capabilities to solve for the aggregation of strong localized sources and diffuse sources (which do not generate clear patterns locally) over large scales with moderate resolution observations (e.g. from GOSAT and now TROPOMI), and the analysis of moderate but smooth variations across continents and countries (see the introduction of the article for regional and global studies of CH4 emission mapping). The integration of CH4 signal through atmospheric mixing often provides good capabilities to monitor budgets of emissions over large regions while limiting those for attributing such an overall emission budgets at finer spatial scales, and in particular at the model spatial resolution, and this phenomenon is exacerbated by the limitations of the spatial and temporal coverage of the satellite data. The statistical inversion noise (combining the observation and model errors) enhance the difficulties to retrieve robust emission estimates at the horizontal resolution of the transport model.

Also, sources with individual facility-level methane emissions greater than approximately 10 t/h are estimated to contribute 8 %–12 % of total global oil and gas production methane emissions (Lauvaux et al., 2022), but have larger contributions locally (Cusworth et al., 2026). Point sources estimates are thus so far limited in their potential to assess national to regional scale budgets, with current observing high resolution capabilities (e.g., Tanager, GHGSat, Emit, EnMAP); dedicated studies are ongoing to find ways to upscale point source estimates to larger scales in a relevant way but these are the work of the upcoming years for the community and will likely justify a large number of dedicated publications. Inversions focus on the smaller emission sources, that account for a large proportion of anthropogenic emissions (Williams et al., 2025), as well as diffuse anthropogenic and biogenic sources. In particular, it makes sense for Europe which has only a few basins with large localized sources (e.g., cf CAMS Hotspot Explorer; https://apps.atmosphere.copernicus.eu/methane-explorer/), whereas most CH4 signals are diffuse and coming from various sources.

For these reasons, both point sources and regional/global inversions are relevant and complementary in their use of the information content of satellite observations. Our publication focuses on the use of satellite images in regional/global inversions and assess limitations specific to regional/global inversions, that are not necessarily critical in hotspot studies. Indeed, retrieval differences will have a lesser impact when studying plumes with very strong signals, whereas in inversions, gradients are smaller and even small observation deviation critically impact inversion results, which is the point of our study.

There is on-going effort to bridge the gap between local point-source estimates to regional and meso-scale inversions (Naus et al., 2023, Visser et al., 2025). The study of Santaren et al. (2021) explores how « atmospheric inversions may have the potential to exploit spatial variations in XCO2 fields to quantify regional to national budgets of more diffuse sources or of all types of sources, even when the overlapping of several plumes prevents the quantification of the emissions from individual cities and point sources ». This is still a large open topic, but goes beyond the scope of this article.

Regarding the comparison with known emissions, gridded inventories such as the one we use as a prior estimate of the anthropogenic emissions (EDGARv8) represent state of the art compilation of "medium and large sources", ingesting as much local information as possible for such a compilation and mapping efforts. This often results in the application of average emission factors and uncertain data on local activites when not only a rough spatial disaggregation of national totals with activity proxies. Local gridded inventories would bring more robust information on local sources but would suffer from the same types of limitations, and (to our knowledge) there is no systematic robust estimate of local sources based on observations over areas that would exceed the scale of our model grid cells (approximately 50km x 50km). Therefore, we hardly see a better confrontation to other emission estimates at our inversion scale than that to gridded inventories such as EDGAR (used in this study) or other similar inventories, which is done in the comparison between prior and posterior emissions.

2. A scatterplot of grid-by-grid, time-by-time is required.If this can be analyzed country-by-country or emissions type-by-type, then this is better, but not required. We need this information to understand if the retrievals are real at the retrieval level, or if the averages make sense for the wrong reasons. The combination of 1 and 2 will ensure that your day-by-day, grid-by-grid emissions are reasonable.

We provide scatter plots of posterior emissions against prior emissions and other posteriors, in the matrix of Figure 3. Each scatter point represent a given date and pixel. This can be used to understand the distributions relatively to others. A second plot is made by aggregating the country yearly budgets and comparing posterior and prior budgets (Figure 4). It shows better consistency at the country scale, yet some differences remain: highlighting the fact that results from the inversions of the TROPOMI products are not fully consistent even on aggregated scales. This is exactly the point of our study to express reserves on using TROPOMI data as such in regional inversion, as has already been published in various papers, including very recently in ACP (e.g., Hancock et al., 2025). We recommend further investigation and caution on regional inversions.

These plots have been added in the Supplementary material. References to the figures and the corresponding discussion has been added to the main text. We have also clarified the scope of the study, and the relevance and the limits of the attribution of posterior emissions to budgets across the different scales.

3. A specific analysis of this should be done in space and time, to ensure that the retrievals are robust over all times of the year. Please check.

The space and time analysis is presented through the seasonal maps of Figures S1, S2 (in the Supplementary material) as well as with the time series of Figure 10c. Given the fact that yearly and country-scale emission totals show already substantial differences across the inversions, we consider that it is not relevant for the conclusions of this study to show detailed analyses at both higher temporal and spatial resolutions at the same time. Given the large number of plots that would be required, we prefer not to extend even more the (already long) article.

4. If they are not reasonable in space and time, and their sum is reasonable, you still cannot do attribution. This is the point of such an inventory. We want to know from a quantitative perspective, if attribution is possible or not. Based on the uncertainty of the emissions from 1-3, please quantify how reliable the product is in terms of attribution.

In our study, we compare posterior estimates of the emissions to the a-priori information coming from bottom-up (BU) inventories. The assimilation of satellite observations of XCH4 provide complementary signal for the mapping of emissions at a 0.5° resolution, as mentioned in the previous responses. In this process, aggregating spatially and temporally the emissions decrease the uncertainty on these emissions.

Previous studies have assessed the finest time and space resolution at which information can be retrieved from the available atmospheric data into the emission space (e.g., Pison et al., 2018). Similarly, studies with OSSEs on point sources (Naus et al., 2023) showed that « daily estimated (point source) emissions have large errors (50%), but these errors mostly average out on monthly and annual scales of 25 and 15%, respectively ». Following this idea, the attribution of emissions can be difficult at the pixel-scale, but still relevant at coarser scales (country, region). This is what we do in the study: we discuss the consistency of top-down emission results at different scales, using different products retrieved from the same L0 measurements. Our work shows a lack of consistency of the emissions at country-scale, therefore comparing products at a finer (i.e. pixel-) scale, and attributing emissions at a finer scale, are beyond the scope of our study and not relevant.

References:

Cusworth, D.H., Bon, D.M., Varon, D.J. et al. Duration of super-emitting oil and gas methane sources. Nat Commun 17, 2011 (2026). https://doi.org/10.1038/s41467-026-68804-7

Hancock, S. E., Jacob, D. J., Chen, Z., Nesser, H., Davitt, A., Varon, D. J., Sulprizio, M. P., Balasus, N., Estrada, L. A., Cazorla, M., Dawidowski, L., Diez, S., East, J. D., Penn, E., Randles, C. A., Worden, J., Aben, I., Parker, R. J., and Maasakkers, J. D.: Satellite quantification of methane emissions from South American countries: a high-resolution inversion of TROPOMI and GOSAT observations, Atmos. Chem. Phys., 25, 797–817, https://doi.org/10.5194/acp-25-797-2025, 2025

T. Lauvaux et al., Global assessment of oil and gas methane ultra-emitters. Science375,557-561(2022). DOI: 10.1126/science.abj4351

S. Naus, J. D. Maasakkers, R. Gautam, M. Omara, R. Stikker, A. K. Veenstra, B. Nathan, I. Irakulis-Loitxate, L. Guanter, S. Pandey, M. Girard, A. Lorente, T. Borsdorff, and I. Aben : Assessing the Relative Importance of Satellite-Detected Methane Superemitters in Quantifying Total Emissions for Oil and Gas Production Areas in Algeria, Environmental Science & Technology 2023 57 (48), 19545-19556 , DOI: 10.1021/acs.est.3c04746

Pison, I., Berchet, A., Saunois, M., Bousquet, P., Broquet, G., Conil, S., Delmotte, M., Ganesan, A., Laurent, O., Martin, D., O'Doherty, S., Ramonet, M., Spain, T.G., Vermeulen, A., Yver Kwok, C.: How a European network may help with estimating methane emissions on the French national scale, Atmos. Chem. Phys., 18, 3779–3798, https://doi.org/10.5194/acp-18-3779-2018, 2018.

Santaren, D., Broquet, G., Bréon, F.-M., Chevallier, F., Siméoni, D., Zheng, B., and Ciais, P.: A local- to national-scale inverse modeling system to assess the potential of spaceborne CO2 measurements for the monitoring of anthropogenic emissions, Atmos. Meas. Tech., 14, 403–433, https://doi.org/10.5194/amt-14-403-2021, 2021.

A. Visser, C. Schooling, N. Bousserez, P. Palmer: Multi-scale global IFS inversion outputs (2021) with assimilated posterior emissions from hot-spots. CORSO project report D2.7, 2025, https://www.corso-project.eu/sites/default/files/2026-01/CORSO-D2-7_v1.1.pdf

Williams, J. P., Omara, M., Himmelberger, A., Zavala-Araiza, D., MacKay, K., Benmergui, J., Sargent, M., Wofsy, S. C., Hamburg, S. P., and Gautam, R.: Small emission sources in aggregate disproportionately account for a large majority of total methane emissions from the US oil and gas sector, Atmos. Chem. Phys., 25, 1513–1532, https://doi.org/10.5194/acp-25-1513-2025, 2025.

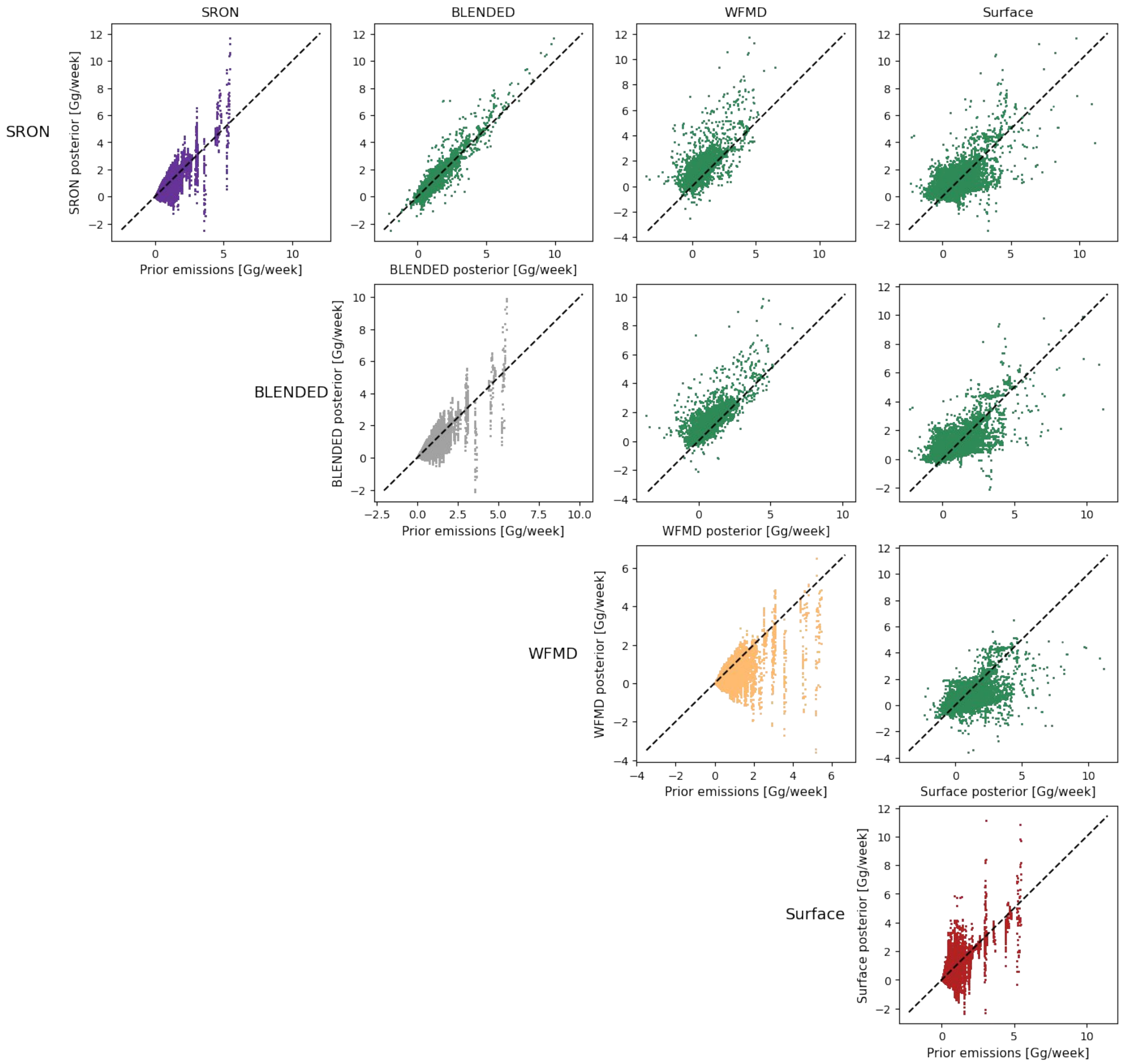

Figure 3: Scatter plots of emissions in the different inversions. The diagonal plots show posterior against prior emissions. The non-diagonal plots show a comparison of the posterior emissions. Each scatter point represents a pixel of 0.5°x0.5° and one week, i.e. One element of the control vector. The dashed line is the 1:1 line.

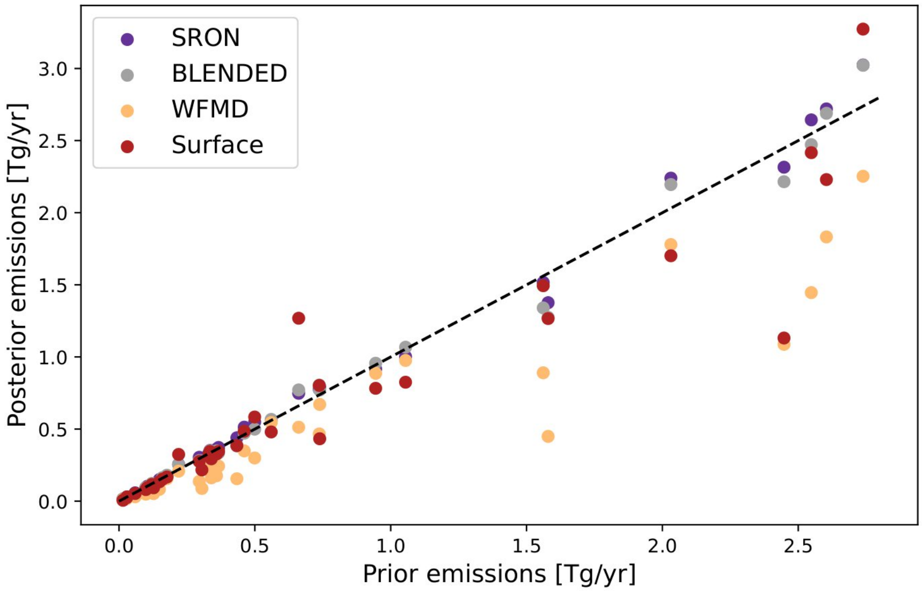

Figure 4: Scatter plot of the yearly and country averaged CH4 prior and posterior emissions, across the different inversions. Each scatter point represents a country.

5. Please follow-up on the response to the fifth question of RC4 for Response-Version 3.

(1) Is the scaling factor for all 8 grids the same across each different posterior? Or is the total amount scaled differently?

If we understand correctly, you ask whether the relative contributions of the 8 pixels in the area vary across posteriors, or if it remains constant and only the total of the emissions varies.

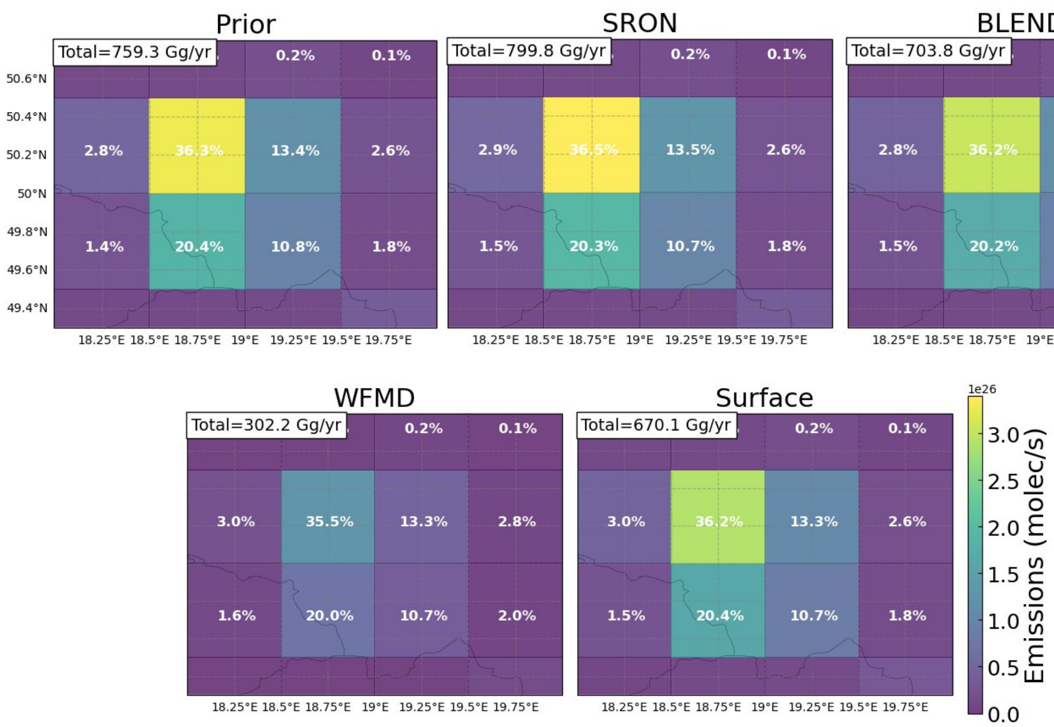

Figure 5: Maps of prior and posterior CH4 emissions for the 4 inversions, over the USCB region studied in Tu et al. (2022), in 2019. Emissions are in molec/s. The contribution of each pixel to the total is written in white.

Figure 5 shows the values of the total prior/posterior emissions and their relative contribution for each concerned grid cells (white figures, in % of the total). The changes in the relative contributions are very small, it is mainly the total over the region that is scaled. It is consistent with what has been said before: the inversions are not designed to attribute very local emissions, but to monitor emission budgets over large regions.

(2) how spatially does this map compare with Tu et al., 2022?

Following what has been said in the response to the first 4 questions, it is not relevant to compare results of our posterior emissions to the results of Tu et al. (2022) at the pixel-level. Given the coarse resolution at this scale, we do not expect the inversion to attribute correctly the emissions from one point-source to the other. Emissions can be compared at the scale of the entire basin to an inventory of the emissions in the area, but not at the scale of point-sources.

Also, as mentioned in the previous response, the study of Tu et al. (2022) has been performed in the year 2021, and extrapolating the emissions to another year (2019 in our case) may introduce errors; indeed emissions in Eastern European countries often vary significantly from one year to the next one, as reported in Romania by Stravapoulou et al. (2023) or Kulhmann et al. (2025). Moreover, the prior estimates of emissions that we use (anthropogenic emissions from EDGAR v8) are far higher than the estimates of Tu et al. As we perform Bayesian inversions, that computes the most probable state given the a-priori knowledge, it is not relevant to compare too deeply the results in the case of large differences in the priors. What we can say is that the total emissions are decreased to values closer to the estimates of Tu et al.

References:

Kuhlmann, G., Stavropoulou, F., Schwietzke, S., Zavala-Araiza, D., Thorpe, A., Hueni, A., Emmenegger, L., Calcan, A., Röckmann, T., and Brunner, D.: Evidence of successful methane mitigation in one of Europe's most important oil production region, Atmos. Chem. Phys., 25, 5371–5385, https://doi.org/10.5194/acp-25-5371-2025, 2025.

Stavropoulou, F., Vinković, K., Kers, B., de Vries, M., van Heuven, S., Korbeń, P., Schmidt, M., Wietzel, J., Jagoda, P., Necki, J. M., Bartyzel, J., Maazallahi, H., Menoud, M., van der Veen, C., Walter, S., Tuzson, B., Ravelid, J., Morales, R. P., Emmenegger, L., Brunner, D., Steiner, M., Hensen, A., Velzeboer, I., van den Bulk, P., Denier van der Gon, H., Delre, A., Edjabou, M. E., Scheutz, C., Corbu, M., Iancu, S., Moaca, D., Scarlat, A., Tudor, A., Vizireanu, I., Calcan, A., Ardelean, M., Ghemulet, S., Pana, A., Constantinescu, A., Cusa, L., Nica, A., Baciu, C., Pop, C., Radovici, A., Mereuta, A., Stefanie, H., Dandocsi, A., Hermans, B., Schwietzke, S., Zavala-Araiza, D., Chen, H., and Röckmann, T.: High potential for CH4 emission mitigation from oil infrastructure in one of EU's major production regions, Atmos. Chem. Phys., 23, 10399–10412, https://doi.org/10.5194/acp-23-10399-2023, 2023.

Author response v5 - 04/06/2026

Response to the Editor

We thank the editor for his important feedbacks. Our author responses are shown in blue.

1) Negative emissions in your results: in your response (Figure 3) and the supplement (Figure S5), scatter plots of posterior emissions show both negative emissions on many grid cells over many weekly time steps. It also seems to show a small number of very high emissions values in some cases over a small number of grid cells over a small number of weekly time steps. You have cited (Zheng et al., 2026 ACP) which demonstrates that physically unrealistic negative emissions arise from the nonlinear propagation of noise and should be filtered out. It is also consistent with the small number of very high emissions points observed. Please clarify:

- Are the values shown in Figure 3 after applying the noise-threshold filter described in Zheng et al., 2026?
- If yes, why do negative values persist?
- If no, please apply the filter and show the filtered results. Alternatively, explicitly state that you have not applied such filtering and discuss the implications for your emission estimates.

Some negative values are already present in the prior dataset, mostly located in two blobs, in the north-western part and in the northern part of the domain (Figure 1, first map). These pixels (6.4% of the pixels) correspond to negative fluxes in the ocean climatology input emissions and have very low absolute magnitude. Other areas with soil sinks have negative prior values, also with low magnitude. They do not significantly contribute to total emissions or European budgets.

These negative emissions can also be seen in the posterior in the inversions. However, as you mention, there are additional negative emissions in the results. The emissions are assumed to be normally distributed, thus the optimization of the control vector allows negative values.

For the SRON, BLENDED, WFMD and Surface inversions, there are respectively 6.6%, 6.4%, 8.2% and 7.1% of the posterior control vector values that are negative (one control vector value = emissions of a 0.5°x0.5° pixel for one week). The additional negative emissions are located in Italy, Central Europe and Northern Africa (Figure 1).

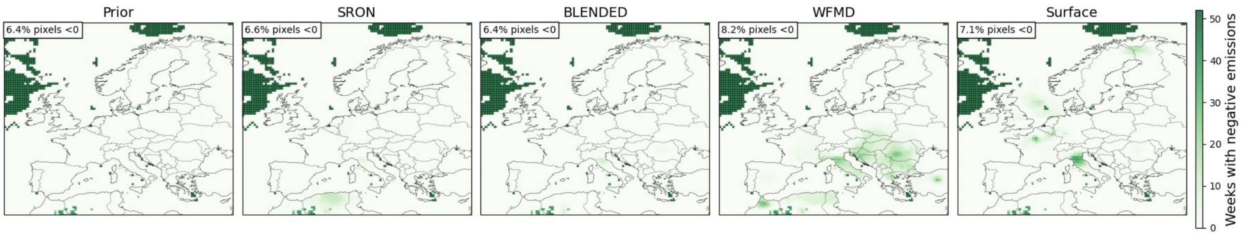

Figure 1: Number of weeks with negative emissions, for all 0.5° pixel, in the prior and posteriors.

As you recommend, we applied the noise-threshold filter described in Zheng et al. (2026). As our inversions are Bayesian inversions (a difference with the cited work), the values that are filtered out in the posterior emissions are replaced by the corresponding prior value. Also, negative emissions corresponding to negative prior emissions (reported natural sinks) are not filtered out. Most strong negative values are indeed filtered out (Figure 2), especially those in the areas mentioned in the previous paragraph (Figure 3). The rest of the distribution is similar, for the four inversions.

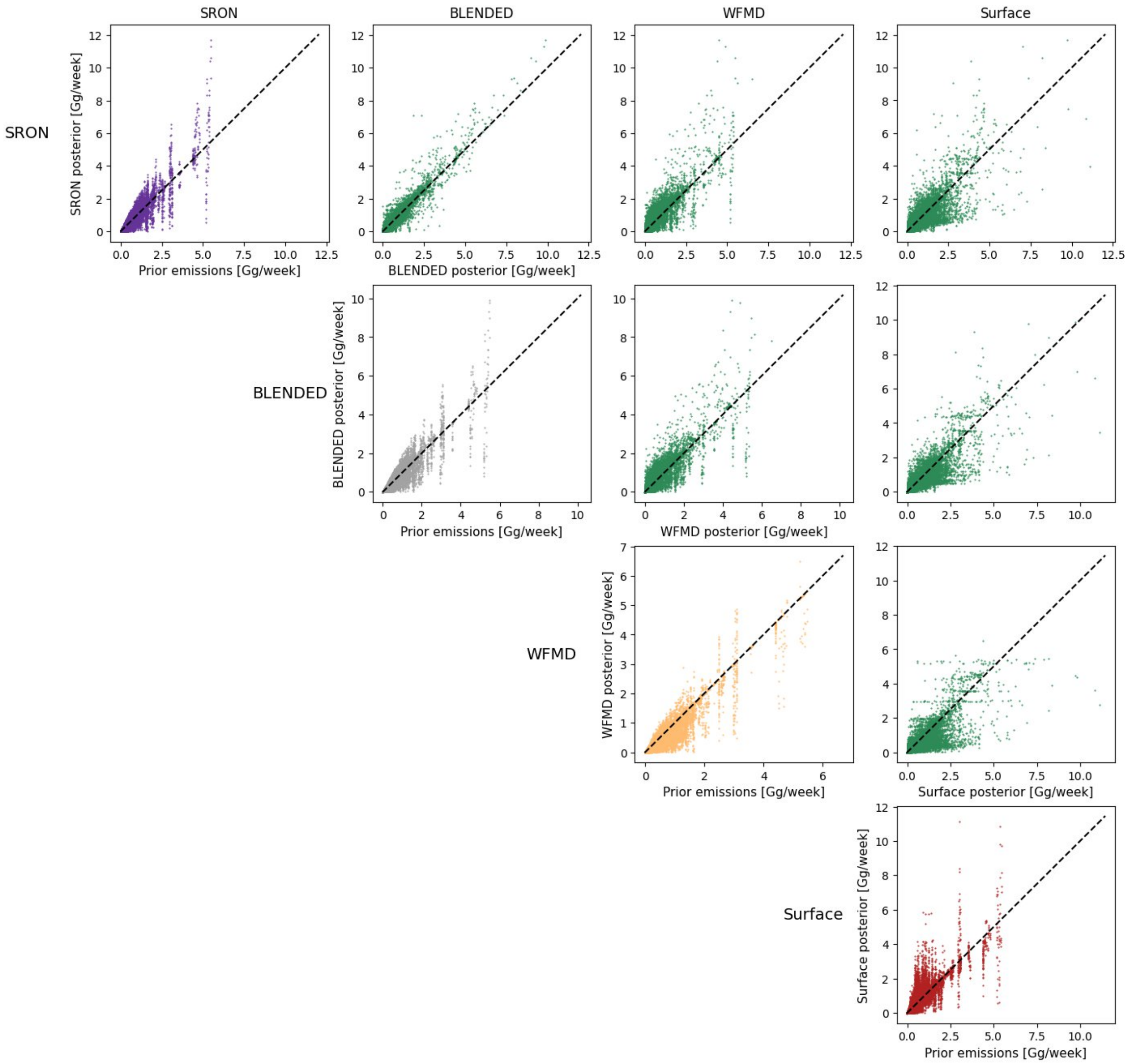

Figure 2: Same as Figure 3 of the previous response, but after applying the noise-threshold filter. The diagonal plots show scatter plots of posterior against prior emissions. The non-diagonal plots show a comparison of the posterior emissions. The dashed line is the 1:1 line.

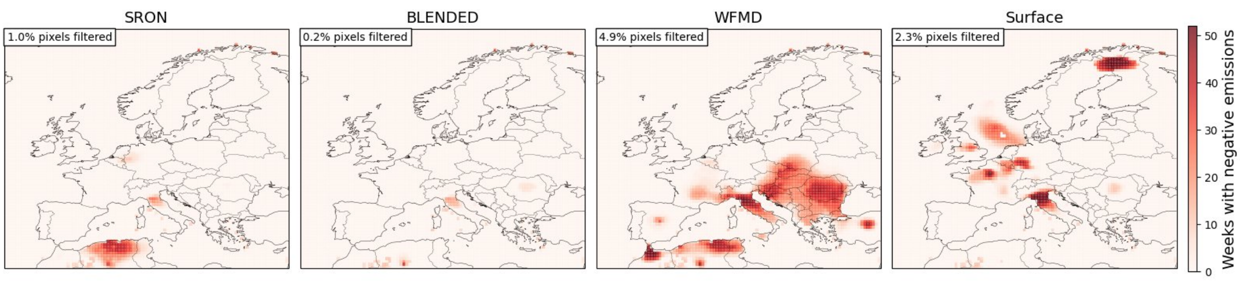

Figure 3: Number of weeks with negative values, for each 0.5°x0.5° pixel, in the prior and the posterior emissions.

The impact of the filter on the CH_4 European emission estimates in 2019 is shown in Table 1. As negative values are removed and replaced with the corresponding prior values, the budget estimates are increased in all the inversions, and are closer to the prior total (except for SRON, for which the estimate is slightly higher than the prior). In particular, the strong negative increments of WFMD in Italy and Central Europe are reduced, leading to a higher total of emissions. In practice, applying this filter gives more confidence to the prior.

When applying the filter, the conclusions of the study remain valid: building country-scale budgets with the emissions estimated from regional inversions of TROPOMI XCH_4 data should be done with caution, as the estimated emissions show significant relative differences across products, in the total and country budgets as well as in the spatial and temporal distributions of emissions.

Inversion	Posterior, without the filter (Tg/yr)	Posterior, with the filter (Tg/yr)
SRON	25.7	26.0
BLENDED	25.0	25.2
WFMD	16.9	21.2
Surface	23.0	25.1

Table 1: European CH_4 budget for 2019, without and with the noise-threshold filter described in Zheng et al. (2026). The total of prior emissions is 25.2 Tg/yr.

Given that the conclusions are not changed, and to remain in the framework of Bayesian inversions, we choose not to apply the filter. Indeed, truncating gaussian distributions in this framework require accurate statistical sampling of the new estimate of the analysis after truncating (Lauvernet et al., 2009), going beyond the variational inversions performed in this study. Though, work has to be done to treat the negative values: either by adapting rigorously the noise-threshold filter to the framework of Bayesian inversions, or by using log-normal distributions for the emissions to ensure positive fluxes (while optimizing sinks separately). The latter and more direct approach, mentioned in Zheng et al. (2026), has been used in recent works (Bergamaschi et al., 2022 ; Hancock et al., 2025). To mention this issue, a discussion on negative emissions has been added to the text (Section 3.3).

2) Spatial scale inconsistency - "not point sources" but still attributing to basins You state that your inversion is designed for basin scale and large sources, not point sources, and that you cannot validate at the native TROPOMI resolution. However, your model grid is 0.5° (approx. 50 km). At this scale, a single grid cell in regions like the Upper Silesian Coal Basin in your region, or many other coal basins around the world (such as the Changzhi Basin) contains a mixture of coal mines, industrial facilities, and natural areas. Therefore, claiming that your results are "basin scale" does not relieve you from the need to demonstrate that your emissions are physically plausible at the grid cell level, especially when you compare to independent local studies (e.g., Tu et al., 2022). Remember that TROPOMI spatial scale (5.5km x 7km approx.) is still not point source scale, yet is far finer than the chosen model scale. Please add a clear statement in the manuscript (and, if necessary, in the response) that:

- Your inversion cannot attribute emissions to specific point sources or to individual source sectors (e.g., coal vs. agriculture) within a 0.5° grid cell.
- Any comparison with local studies (e.g., Tu et al., 2022) should be interpreted with caution because of the scale mismatch.

The statements have been added to the manuscript (Section 3.3 and Conclusion). We confirm that the inversion can not disentangle source sectors within a 0.5° grid cell, as well as individual point sources as mentioned in the previous response. The expression « basin scale » may have been poorly used in our response: we claim that it is not relevant to compare directly the emissions at the grid cell level, this is why we compare national budgets and average maps in the article.

The comparisons with local studies have been made to address the questions of a reviewer: we definitely agree that they should be interpreted with caution. We emphasize this point in the clarifications that have been added to the text.

3) Day-by day, grid by grid validation: a reviewer requested a scatterplot of modeled vs. observed XCH4. Please provide a representative example (e.g., for a few weeks in a key region) showing the correlation between observed and simulated XCH4 after the inversion, aggregated to your 0.5° model grid. This will help readers assess whether the inversion is fitting the data in a physically reasonable way.

> Additionally, for a small subset of grid cells (e.g., 5-10 cells covering different land uses), please report the standard deviation (or 10th-90th percentile range) of the native TROPOMI pixels that contribute to each 0.5° average. This will indicate whether the averaging hides significant spatial heterogeneity that could bias the inversion.

The scatter plots of observed and simulated XCH4 before and after the inversion are shown in Figure 4 for all observations, aggregated to the 0.5° model grid. The fit between modeled and observed XCH4 is improved, as shows the increase of the correlation coefficient R^2 for all inversions . When looking at a specific region (2°-10° longitude, 48°-54° latitude, covering Germany, the Netherlands, Belgium, Luxembourg and eastern France) where the initial fit is poor, the fit of the data after the inversion seems physically reasonable, with a substantial increase of R^2 (Figure 5). Figure 4 has been added to the supplementary material, complementing the maps of Figure 9 in the manuscript.

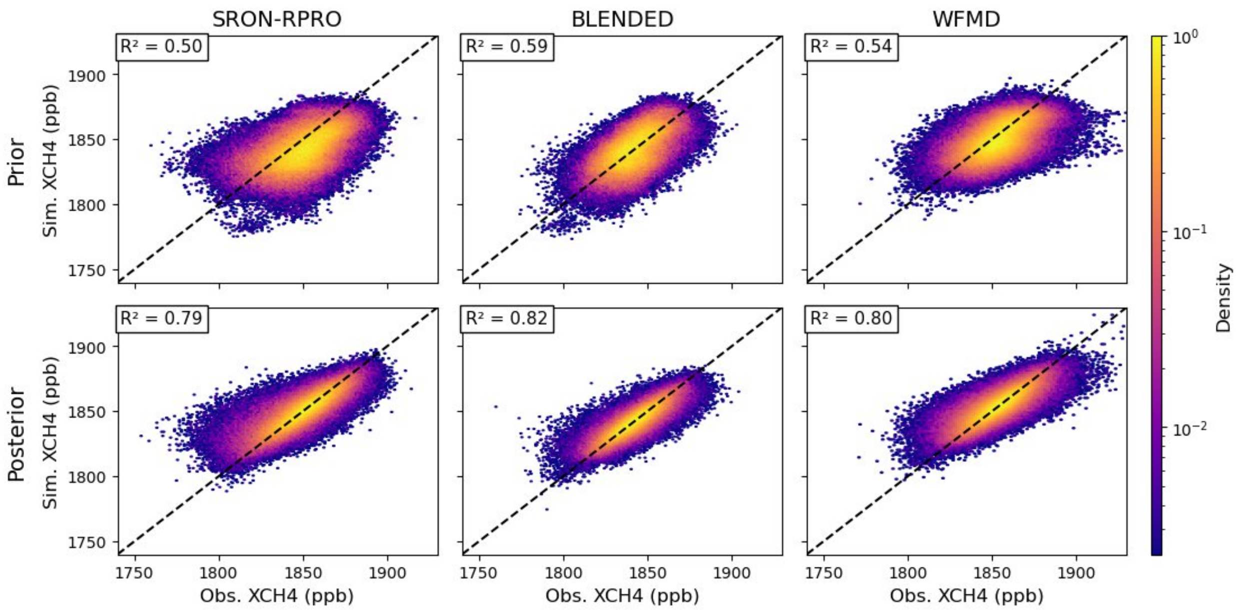

Figure 4: Hexagonal binning plots of the modeled and observed XCH_4 for the satellite inversions, using prior emissions (first row) and posterior emissions (second row). R^2 is the correlation coefficient. The dashed black line is the 1:1 line.

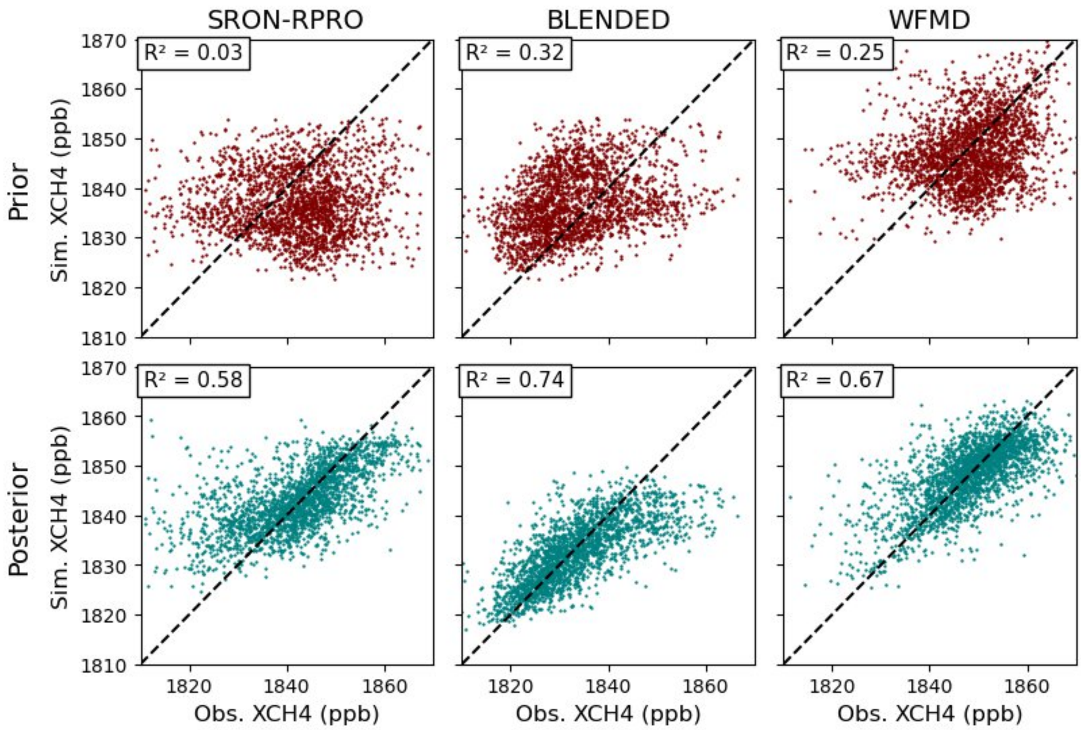

Figure 5: Scatter plots of the modeled and observed XCH_4 for the satellite inversions, using prior emissions (first row) and posterior emissions (second row), in the region covering 2°-10° longitude and 48°-54° latitude.

Regarding the spatial heterogeneity, the standard deviation of TROPOMI XCH4 in the 0.5°x0.5° pixel is shown for a few pixels in Figure 6. The selected pixels are shown on the map, and for all products the deviation are in the range 10 to 25 ppb, which is of the order of magnitude of the observation errors. When considering all pixels, the deviations mostly range between 0 and 25 ppb, with a maximum of pixels below (SRON, BLENDED) or around (WFMD) 10 ppb (Figure 7).

In the inversion, observations are assimilated individually: the inversion is not directly affected by aggregation effects, though the system has to fit multiple observation values that are in the same model spatio-temporal grid cell. This is an indirect effect of the spatial heterogeneity. Accounting for this heterogeneity could be done through the calculation of the aggregated observation error if super-observations were used (Chen et al., 2023), which is not the case in our study.

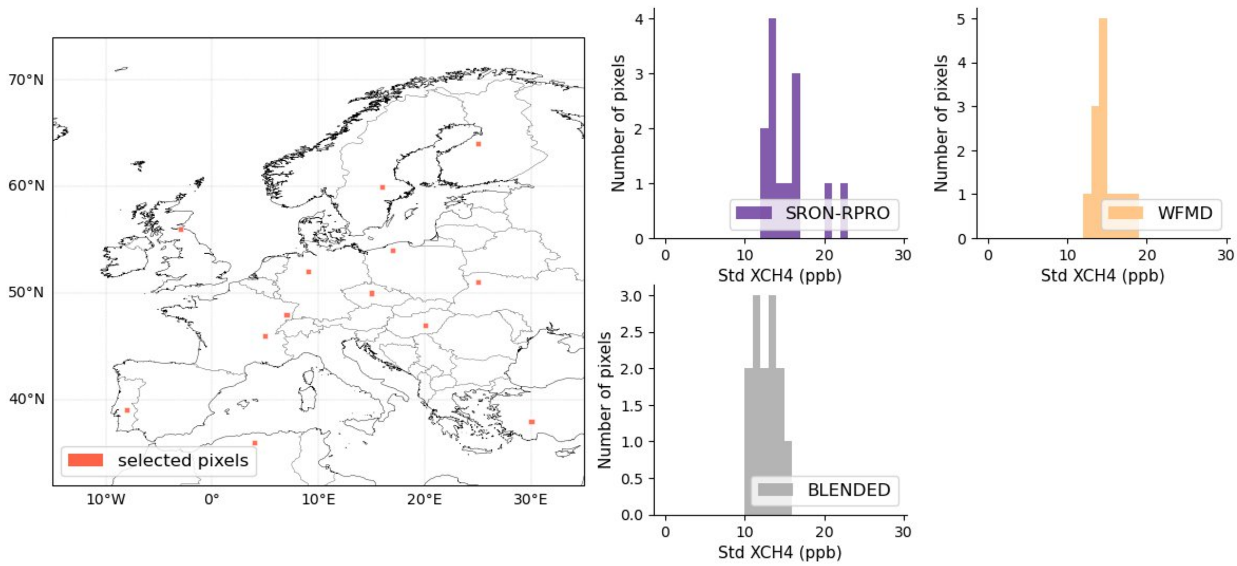

Figure 6: Standard deviations of the XCH4 observations contributing to the 0.5° averages, for all TROPOMI products and for the selected pixels pictured on the map.

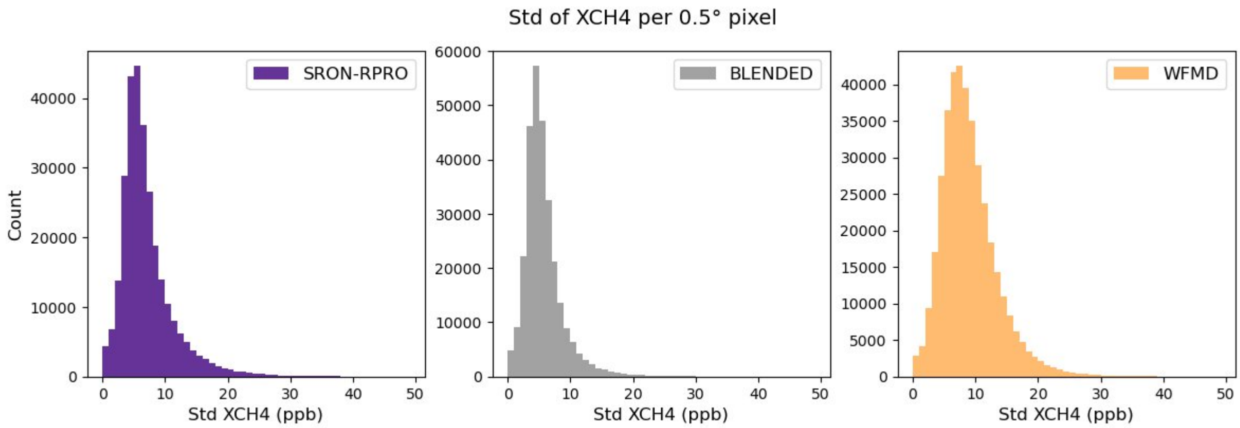

Figure 7: Distributions of the standard deviations of the XCH4 observations contributing to the 0.5° averages, for all TROPOMI products and all pixels.

References:

Bergamaschi, P., Segers, A., Brunner, D., Haussaire, J.-M., Henne, S., Ramonet, M., Arnold, T., Biermann, T., Chen, H., Conil, S., Delmotte, M., Forster, G., Frumau, A., Kubistin, D., Lan, X., Leuenberger, M., Lindauer, M., Lopez, M., Manca, G., Müller-Williams, J., O'Doherty, S., Scheeren, B., Steinbacher, M., Trisolino, P., Vítková, G., and Yver Kwok, C.: High-resolution inverse modelling of European CH4 emissions using the novel FLEXPART-COSMO TM5 4DVAR inverse modelling system, Atmos. Chem. Phys., 22, 13243–13268, https://doi.org/10.5194/acp-22-13243-2022, 2022.

Chen, Z., Jacob, D. J., Gautam, R., Omara, M., Stavins, R. N., Stowe, R. C., Nesser, H., Sulprizio, M. P., Lorente, A., Varon, D. J., Lu, X., Shen, L., Qu, Z., Pendergrass, D. C., and Hancock, S.: Satellite quantification of methane emissions and oil–gas methane intensities from individual countries in the Middle East and North Africa: implications for climate action, Atmos. Chem. Phys., 23, 5945–5967, https://doi.org/10.5194/acp-23-5945-2023, 2023.

Hancock, S. E., Jacob, D. J., Chen, Z., Nesser, H., Davitt, A., Varon, D. J., Sulprizio, M. P., Balasus, N., Estrada, L. A., Cazorla, M., Dawidowski, L., Diez, S., East, J. D., Penn, E., Randles, C. A., Worden, J., Aben, I., Parker, R. J., and Maasakkers, J. D.: Satellite quantification of methane emissions from South American countries: a high-resolution inversion of TROPOMI and GOSAT observations, Atmos. Chem. Phys., 25, 797–817, https://doi.org/10.5194/acp-25-797-2025, 2025.

Lauvernet, C., Brankart J.-M., Castruccio, F., Broquet, G., Brasseur, P., and Verron, J.: A truncated Gaussian filter for data assimilation with inequality constraints: Application to the hydrostatic stability condition in ocean models, Ocean Modelling, 27, 1–2, 1-17, https://doi.org/10.1016/j.ocemod.2008.10.007, 2009.