This paper uses a suite of Observing System Simulation Experiments (OSSEs) to explore the capability of the global surface flask monitoring network and various satellite instruments to constrain surface fluxes of methane (CH_4). This technique is well-established and has been used previously in similar studies, but the authors here do a good job of updating such work and expanding the analysis in a nice way. The use of small Monte Carlo runs for each inversion is a good way to probe the posterior uncertainties of the results, although the ensembles are of **fairly limited size.**

The diagnostics are well-chosen and defined in an intuitive way, on the whole. The study's findings imply that observation density alone is not sufficient to improve inversion performance, particularly if it comes at a trade-off with observation uncertainty. It also reinforces that a good surface flask network is of vital importance for top-down emission monitoring and must be maintained and, preferably, expanded significantly in order that we can better understand fluxes of species like CH_4 in the future.

Overall, the paper is well-written, structured properly and has useful and well-designed figures. In my opinion it will serve as a useful reference for those attempting to do top-down estimations of CH_4 fluxes in the future and should be published after the following minor suggestions have been incorporated.

Response

We appreciate your feedback on our work. Your constructive comments have helped us improve the presentation of our work in terms of scientific quality and structure.

Comments:

Figure 1: Please consider using a colour-blind friendly colour table for this figure (and others in the Appendix).The rainbow colourbar is not reader-friendly for such figures for a number of reasons. See https://blogs.egu.eu/divisions/gd/2017/08/23/the-rainbow-colour-map/

Response:

We have taken this into account; it is important that the article be as reader-friendly as possible.

Line 438: How was the ensemble size of 10 members arrived at? Are you satisfied that it could represent the full uncertainty spectrum accurately? Please justify here.

Response:

The ensemble size of 10 members was primarily constrained by computational cost. In this study, 14 different configurations were explored (including the simple system and the different combined-system configurations), resulting in a total of 140 inversion runs. Each inversion required access to 10–14 CPUs together with GPU resources, and a single ensemble member typically required approximately one week of computation time to complete. On average, each inversion reached around 60 iterations under a strict convergence criterion of 0.001 EPSG (precision of the stopping criterion based on the norm of the gradient).

We acknowledge that a larger ensemble would likely provide a more exhaustive characterization of the uncertainty and potentially improve the robustness of the statistical estimates. However, within the available computational resources, an ensemble size of 10 members represented a practical compromise between computational feasibility and uncertainty exploration. Despite this limitation, the ensemble spread was sufficient to identify the main variability patterns and assess the sensitivity of the inversion results across the tested configurations.

Line 435: Did you consider randomly perturbing the synthetic observations within their precision to represent satellite 'noise'? Wouldn't this more accurately represent the satellite system in the OSSE and reduce the likelihood that you're overestimating the error reduction? What is the impact of this decision on your results – please discuss at some point.

Response:

Thank you for this comment. We agree that perturbing the synthetic observations according to their measurement uncertainties would provide a more realistic representation of the observing systems and would likely lead to lower uncertainty reductions.

Our choice to perturb only the prior fluxes was intentional and reflects the highly idealized nature of the OSSE framework. The objective of this study was not to reproduce the full complexity of real-world inversions, but rather to establish a controlled framework in which the different observing systems could be compared on an equal basis. In this sense, the results should be interpreted as an estimate of the maximum potential performance of each observing system under ideal conditions.

Introducing random perturbations in the observations would constitute an interesting extension and would likely reduce the magnitude of the estimated uncertainty reductions. However, it would also introduce an additional source of variability, making it more difficult to isolate the relative contribution of the observing systems themselves. We have clarified this point in the manuscript and emphasize that our results represent an optimistic benchmark rather than a fully realistic inversion scenario.

In revised text, we added:

In addition, the synthetic observations were not randomly perturbed according to their measurement precision. This choice was made to facilitate a direct comparison of the observing systems under idealized conditions and therefore the reported uncertainty reductions should be interpreted as an optimistic estimate of their potential performance.

Line 453: Please briefly explain why the monthly averaging is done in log space before returning to real-space values.

Response:

The monthly averaging was performed in log space in order to reduce the influence of extreme monthly values and to provide a more balanced aggregation of the relative error reduction metric. Since the quantity inside the logarithm is a ratio between posterior and prior misfits relative to the truth, its distribution can span several orders of magnitude across months and regions. Averaging directly in real space would therefore allow a small number of extreme cases to dominate the annual metric.

By averaging in log space, the metric effectively treats improvements and degradations in a more symmetric multiplicative sense and reduces sensitivity to outliers. The final transformation back to real space in Eq. 9 preserves interpretability of the metric in percentage terms.

Lines 461 – 463, and Section 4.1: The author's discussion of the M metric can be confusing in places, in my opinion. For example, lines 462-463 state "…whereas positive values indicate that the posterior fluxes fail to improve upon the prior in reproducing the truth (**darker blue in Figure 6)**." However, darker blue colours relate to negative values and red colours relate to positive M values.

Response:

Thank you for pointing this out. You are right: the text was inconsistent with the color scale shown in Figure 6. Negative M values correspond to improved posterior performance (darker blue colors), whereas positive values indicate a degradation relative to the prior (red colors). We have corrected the text in lines 461–463 to remove this confusion and ensure consistency with the figure.

Elsewhere in Section 4.1 and beyond, the M value is referred to as positive when it is in fact negative (see line 509-510 " M values above 20%"), whereas on line 513, similar M values are referred to as "below -20%". Please check these sections and be consistent with your description of M values. This has been fixed in the text

Line 510-514: I'm afraid I don't quite follow your point on regions (such as China) with low sensitivity and high flux densities 'appearing' improved when this is not really the case. Can you clarify here?

Response:

Thank you for this comment. We agree that the original wording was not sufficiently clear and could lead to a misleading interpretation. Our intention was not to suggest that regions such as China are poorly constrained despite showing improved posterior estimates, but rather to highlight that good posterior performance may result not only from strong observational sensitivities but also from the magnitude of the underlying emissions.

We have therefore revised the text as follows:

Some regions, such as China, combine relatively modest observational sensitivity with very high flux densities. In these cases, good posterior performance does not necessarily reflect stronger observational constraints, but rather the fact that large emissions generate a stronger atmospheric signal, making flux perturbations easier to detect.

Line 551: You mention representativeness errors here – am I correct in thinking that you have used a flat 10 ppb value for these errors on all satellite data, based on a 2014 paper relating to SCIAMACHY? Are you confident that this remains a useful assumption and would you expect a more complex model for representativeness error to change your results for the different satellite instruments included here?

Response:

The additional 10 ppb term corresponds to a model representation error added to the observational uncertainty matrix, rather than to the instrumental error of the satellite retrieval itself. Its purpose is to account for unresolved discrepancies between satellite observations and model-simulated concentrations, including transport errors, sub-grid variability, and differences in spatial representativeness.

We acknowledge that more sophisticated, instrument-dependent representation error formulations could be used. However, the constant 10 ppb term was adopted following previous inversion studies and represents approximately 1% of typical concentration values, thereby acting as a conservative regularization to prevent over-weighting satellite observations. Assessing the sensitivity to alternative representation error parameterizations is beyond the scope of the present study.

To make this point more explicit in the revised manuscript, **we reformulated this section as follows:**

For satellite observations, $\mathbf{R}$ includes the reported retrieval uncertainty and an additional 10~ppb term accounting for model representation errors associated with the comparison between satellite observations and model-simulated concentrations. This additional term does not represent instrumental uncertainty, but rather unresolved errors arising from model transport, sub-grid variability, and differences between the observation footprint and the model resolution. Following \citet{locatelli2015sensitivity} and \citet{cressot2014consistency}, a constant value of 10~ppb (approximately 1% of typical observed concentrations) was adopted to avoid over-constraining the inversion with satellite data.

Line 580 – 583: I don't follow this logic. Please clarify? Why does the addition of GOSAT data degrading performance relative to the surface-only inversion having few sites there? Surely this is true in both cases? Apologies for missing the point!

Response:

Thank you for this comment. We agree that our original explanation was not sufficiently justified. In theory, adding observations should either improve the inversion performance or have only a limited impact, rather than systematically degrade it. The decreases observed in Northern and Southern Africa are relatively small (typically within about 10 percentage points) and are therefore difficult to interpret as a meaningful deterioration in performance.

Given that our uncertainty estimates are based on an ensemble of only 10 members, these differences may simply reflect sampling variability within the experiment. A larger ensemble would likely provide a more robust estimate of the uncertainty reduction and help determine whether these small changes are statistically significant. We have therefore revised the text to avoid overinterpreting these modest differences.

Revised text (text added):

Small decreases in M and UR are observed in Northern and Southern Africa when GOSAT observations are added. Given their limited magnitude and the finite size of the ensemble used in this study, these differences should be interpreted with caution and may reflect sampling variability rather than a systematic degradation of inversion performance.

Line 607: Rewrite this statement slightly – it currently implies that both M and UR are 2x GOSAT-NIES and 4x GOSAT-Leic, which I don't think is true.

Revised text

In the 60°N–30°N latitudinal band, GN+S inversions provide the strongest performance, achieving an M value of approximately −40% and a UR exceeding 50%. This UR is approximately twice that obtained with GOSAT–NIES alone and four times that obtained with GOSAT–Leicester alone.

Line 842: In this section you mention the possibility of other combinations of satellites. Why did you choose not to include TROPOMI + surface, TROPOMI + GOSAT or IASI + other data in this manuscript? This would surely have strengthened the results. Although I understand that you might be reluctant to complete new model runs at this stage, I think that discussion of the other available combinations is warranted, at least.

Response:

We agree that exploring additional combinations of observing systems (e.g., TROPOMI + surface, TROPOMI + GOSAT, or IASI combined with other datasets) would provide a more comprehensive assessment of the complementarity between current observing systems and their respective uncertainty reduction capabilities.

In fact, we attempted several additional configurations beyond those presented in the manuscript. However, these experiments encountered significant technical and computational limitations.

In particular, the combined TROPOMI + surface and IASI + surface configurations generated extremely large observation vectors, leading to memory limitations during the inversion procedure, even after applying super-observation strategies to reduce the observational volume. In addition, experiments combining SWIR and TIR satellite products (e.g., GOSAT + IASI) required developments that are not yet fully operational within the CIF platform for this specific multi-satellite configuration framework.

We nevertheless fully agree with the reviewer that including a broader range of combined observing systems would further strengthen the analysis and provide an even more complete comparison of the performances of current observing systems in terms of uncertainty reduction and recovery of prior emissions. We therefore added discussion in the revised manuscript acknowledging these additional possible configurations and their interest for future work.

In this section: 5.6 Satellite Observing Systems: Current Capabilities and Future Perspective, we added :

Additional observing-system combinations were also explored during the development of this study, including satellite + surface assimilations using TROPOMI and IASI, as well as SWIR+TIR satellite combinations such as GOSAT + IASI. However, these experiments encountered important technical and computational limitations. In particular, configurations combining high-density satellite datasets with surface observations generated extremely large observation vectors, leading to memory limitations despite the application of super-observation strategies. Similarly, joint SWIR+TIR assimilations require further developments within the CIF framework to fully support these multi-satellite configurations. Nevertheless, these combined observing systems remain highly promising for future OSSE and inversion studies, as they could provide a more comprehensive assessment of the respective contributions of current and emerging methane observing systems to uncertainty reduction and posterior flux estimation. These configurations represent a natural next step for evaluating the complementarity between observing systems with different spatial coverage, precision, and vertical sensitivity.

Conclusions section: To what extent are your results limited by the fact that you use a single atmospheric model? Would you expect these conclusions to be upheld if an ensemble of models were used?

Response:

The inversion is driven by the Jacobian (or gradient) matrix, which links changes in emissions to changes in the observations. Therefore, the inferred sensitivities depend on the transport model and on how it represents processes such as horizontal and vertical transport, convection, and interhemispheric exchange.

For example, stronger vertical mixing may dilute surface emission signals, while stronger horizontal mixing may increase the footprint of an observation but reduce the ability to distinguish between nearby source regions or sectors. As a result, the magnitude of the uncertainty reductions and regional sensitivities would likely vary across transport models.

That said, we expect the main conclusions regarding the relative performance of the different observing systems to remain qualitatively robust, as they are largely driven by differences in observational coverage and sampling rather than by the transport model itself.

We added in “Scope, sources of uncertainty, and limitations in OSSE” :

An additional limitation is that all experiments rely on a single atmospheric transport model. The inversion sensitivities are determined by the Jacobian matrix linking emissions to observations, which depends on the representation of atmospheric transport processes such as vertical mixing, horizontal transport, convection, and interhemispheric exchange. Different transport models may therefore yield different regional sensitivities and uncertainty reductions. While the quantitative results could vary, we expect the main conclusions regarding the relative performance and complementarity of the observing systems to remain qualitatively robust.

Technical corrections:

Line 13: "OSSEs are conducted over 18 regions…" is misleading. Change to "OSSE results are analysed over…", or similar? replaced by The OSSE results are analyzed across 18 regions ...

Line 17: Should "…reducing uncertainties of more than 30%..." be "…reducing uncertainties by more than 30%..."? change accepted

Line 25 and line 49: mixing ratio -> mixing ratios change accepted

Line 66-67: remove "the" in "… the Japan's GOSAT mission…" change accepted

Line 471: "$UR_{r,}t$ = …". The 't' should be subscript. change accepted

Line 494: "Figures 8" -> "Figure 8" change accepted

Line 608: Fix "Result reflecting…" changed by This highlights ..

Figure 10 caption: Fix "10% was applied 10%". fixed

Line 733: Insert space after "…CO_2)." inserted