Empa
Materials Science and Technology

Reply to the comment of Referee 1

We would like to express our gratitude to the referee for the effort in reviewing our manuscript and for the constructive feedback. We have considered all comments and suggestions and address each point in this response letter, with reviewer comments in blue and author responses in black. We believe that the corresponding revisions to our manuscript will enhance its clarity, accuracy, and overall quality.

Steiner et al. investigated the impact of flow-dependent transport uncertainties on the methane inversions and found that in the OSSE setting that this approach of specifying more realistic errors improves the accuracy of the inversion. The model transport error is an important yet often underdiscussed factor in the inverse modeling, hence the method proposed here is useful for improving error specification in regional-and continental-scale inverse modeling. Below are my comments, mostly suggestions for clarification. The paper is publishable once they are addressed.

L6 and L16: L6 : flow-dependent method improves accuracy, L16: improves precise. It appears that most of the discussion focuses on accuracy rather than precision.

This is correct, and we have adjusted the terminology accordingly so that we now use accuracy instead of precision.

L99-102: the description here is confusing. L99: "we optimized only the emissions, but not the background..". L102:"optimization of background levels together with emissions". Some clarification is needed.

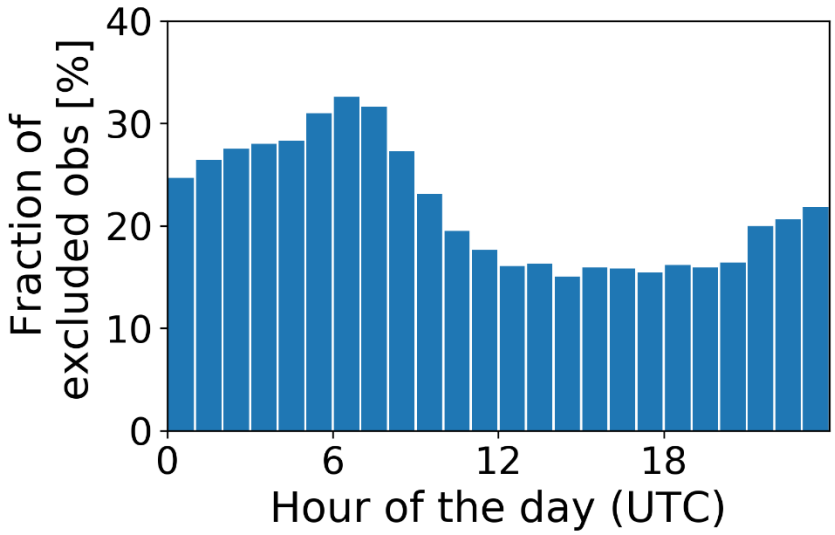
In the idealized setup, we did not optimize the background concentrations because the errors in the background (due to the use of different meteorology and different background concentrations) were part of the artificial transport error. Consistent with this, the ensemble spread caused by the different background concentrations contributed to the overall ensemble spread that determined the mdm. However, in a real data application where non-random biases may occur, it is still useful to optimize the background. This is what we did in our real data application.
We have revised the sentences for clarity as follows:

"In contrast to our previous study, we optimized only the emissions in the idealized setup. We did not optimize the background concentrations because the differences in background CH_4 concentrations introduced by deviating from ''true'' meteorology and using perturbed background CH_4^{bg} concentrations in the driving data are part of the artificially created transport error and contribute to the ensemble spread that determines the mdm. However, if there were systematic biases in background CH_4 in an application with real data, it would still be necessary to optimize background concentrations together with the emissions."

L125-126: (1) In most part of the paper, standard deviation of the ensemble is used (e.g., fig.4). But here "ensemble spread" is used. Are they the same thing? (2) Would this procedure mostly exclude nighttime observations? This information may be useful for readers to assess the difference of this treatment from using daytime only observations.

Yes, in line 125, by "ensemble spread," we meant the standard deviation. We have updated the terminology to "standard deviation" for clarity.

The following figure shows the fraction of excluded observations per hour of the day (from UTC time). In fact, more observations are excluded during the night hours, but also about 15% of daytime observations are excluded. A rough division into day (08 - 20 UTC) and night (all other hours) shows that 1735 day observations and 2514 night observations are excluded. We have added this figure to the manuscript.



L143: What is SPPT? What "observations" is referred to here? Weather observations, or GHG observations? Spell out SST as well.

We have clarified that SPPT stands for "Stochastically Perturbed Parameterization Tendencies" and added this explanation to the manuscript along with the appropriate reference. We have also spelled out SST. In line 143, "observations" referred to weather observations. However, also the GHG observations are perturbed, and we have now included this information in the text.

L170: The equation is not formally written, by mixing the matrix/vector form with the scalar form. Better to write in the matrix/vector form. 1/n (y-Hx)T (HPH+R)^-1 (y-Hx)

Thank you for bringing this to our attention. We have corrected the formula accordingly.

Section 2.6. A lot of discussion regards to flat-terrain sites vs. mountain sites. But no information is provided in Fig. 7 or 12 which sites are classified as flat-terrain or mountainous. It would be good such information can be provided.

We've marked the mountain stations with triangles in Figures 7, 8, and 12 and we note this in Section 2.6 of the manuscript.

L220: Only temporal correlation is considered in this study. However, Fig. 3 indicates that transport model may also cause "flow-dependent" spatial error correlations, by comparing the error patterns in Fig. 3 and the site distribution in Fig. 12. It may make sense to derive this spatial error correlations from the perturbed physics ensemble, and assess the impact of spatial error correlation on the inversion results.

The referee is right that these spatial correlations could also be sampled from the ensemble. However, due to the small number of ensemble members, these correlations are very noisy. One would have to apply a diffusion model, as Lauvaux et al. (2009) did, to extract the significant part of the noisy error. Their diffusion model is based on the method of Pannekoucke and Massart (2008), which estimates the structural properties of the ensemble variability using a diffusion operator based on a local diffusion tensor. While we have considered such an approach, we have decided against implementing it due to the considerable effort involved, especially given our setup covering the large domain of Europe, where spatial distances between stations are often (though not always) large, making temporal correlations much more critical than spatial correlations in our application.

L225-226: For conditioning the R-matrix, what is the reasoning to apply an exponential decay of exp(-deltaT/24h)? Would this defeat the purpose of error correlation fitting performed above?

A detailed rationale for the use of shrinkage methods or smoothing of off-diagonal elements can be found in Ghosh et al. (2021). Spurious correlations arise due to the small ensemble size, resulting in poor conditioning of the R matrix when considering only the sampled covariances in the off-diagonal elements. As described in the manuscript, this leads to instability and produces completely unrealistic values. Therefore, Ghosh et al. (2021) proposed two different approaches. In addition to smoothing by exponential damping of the off-diagonal elements, which we also use, they applied a shrinkage-based regularization of the covariances. Both approaches yielded similarly good results, although they inherently reduce the influence of true correlations to some extent. While correlations with observations in the following hours remain well preserved, correlations

between nighttime observations tend to be significantly attenuated, affecting their reappearance in the following night. Nevertheless, as shown in our synthetic setup, our results still show improvement, allowing us to extract more information from the observations overall. We also tried a longer length scale for the damping factors of 30 hours instead of 24 hours, but this again led to unstable results in the inverse of **HPH+R**.

Fig. 3: the unit of Fig. 3 is kg/kg, which is an uncommon unit for CH4, and is inconsistent with the unit (ppb by volume) used elsewhere in the paper.

We have changed the unit in the figure correspondingly.

Fig. 5: This figure shows interesting information. But I am confused why in panel c and d, the STD(CH4_emis) is non-zero for the perturbed IC/BC only case. If I understand correctly, CH4_emis tracer is not affected by IC/BC but changes with wind fields in the modeling domain. In the perturbed IC/BC only case, the wind fields within the modeling domain are identical across the ensemble member, and, therefore, should result in the same CH4_emis tracer.

"Perturbed IC/BC only" means that the ensemble experienced only perturbed initial and boundary conditions, without additional perturbation by perturbed model physics. Perturbed initial and boundary conditions mean that the driving meteorological data (and the CH_4 field used to drive the background) are perturbed, resulting in each ensemble member having a different meteorology. This leads to differences in the emitted CH_4 tracer. We have revised the text to clarify that perturbed IC/BC refers to the way the ensemble is generated, not just the CH_4 concentration field. Specifically, we have changed the following sentence:

"Each figure presents two scenarios: one with only perturbed initial and boundary conditions (gray), and the other one with both, perturbed model physics and perturbed initial and boundary conditions (IC/BC, red)."

to

"Each figure presents two scenarios: one where the ensemble is generated with perturbed initial and boundary conditions only (IC/BC, gray), and one where in addition to IC/BC also the model physics is perturbed (IC/BC + STTP, red).".

L317: and -> with

Corrected

Section 3.4: It appears that the real-data inversion uses only daytime average data. Is it possible to use hourly data filtered by flow-dependent errors, as shown in Section 3.3?

This is technically possible. However, while the assimilation of hourly data yields better results in OSSE, it is unclear whether this holds true for real measurements. This would require that the ensemble spread accurately reflects the real uncertainties, and that the ensemble simulation does not have substantial biases at night. The latter is probably the more critical issue. Therefore, we refrain from including these results in the publication and limit our demonstration to the idealized setup.

Section 4: This long paragraph is difficult to read. I'd suggest to organize the section into separate paragraphs.

We have divided the Conclusions into individual paragraphs, each with its own title.

L395-403: This interesting discussion should belong to "Results and Discussion" instead of Conclusion.

We have moved most of this discussions to "Results and Discussion", and briefly mentioned it again in the Conclusions.

Reply to the comment of Referee 2

We would like to thank the referee for taking the time to read our manuscript and for the thoughtful feedback. The input was very valuable to enhance the quality and rigor of our work. We have taken all suggestions along with the comments of Referee 1 into account and present our responses below, with reviewer comments in blue and author responses in black.

General comments

The paper describes experiments with a new implementation of the observation-representation-error in an Ensemble Kalman smoother system that is used for emission inversion of methane. The new method uses ensemble simulations to quantify the transport uncertainty at observation sites from the ensemble spread. The ensemble perturbations are driven by a meteorological ensemble and model changes, and therefore time- and flow-depended.

The authors show that the new dynamic formulation of the observation-representation-error (or as called in the paper, 'mdm' or 'model-data-mismatch') is beneficial over an implementation using a static error. It allows for example the use of CH4 observations for every hour of the day rather than afternoon or nighttime averages only as is currently common practice, with improved result. Although running an *a priori* ensemble is not cheap, it seems worthwhile the effort. This is an important conclusions, and the paper is therefore a useful contribution to the field of greenhouse-gas inverse modelling. Could the authors give an indication of the costs (computational, storage, work?) of the proposed method? Compared to regular inversions, are the costs minor or substantial? Could external data suppliers be of use here and make dedicated data available for the purpose of this method?

The computational cost, measured in node-hours, for the a priori ensemble simulation with 10 members (without the flux ensemble with 192 tracers) was about 1.7 times higher than the cost of regular inversions. This means that the total computational cost (prior ensemble + inversion) was 2.7 times the cost of a regular inversion, which is a considerable but not prohibitive increase. We recognize that this important information should be communicated in the manuscript, and have therefore added it to the Methods section (subsection "ICON-ART ensemble simulations").

In terms of storage, the output of the a priori ensemble simulation required about half as much space as the output of the regular inversion. Thus, the storage requirement was 1.5 times the size of a regular inversion.

It is important to emphasize that we did not optimize the a priori ensemble simulation for computational time and memory. We wrote out significantly more variables than were necessary to determine the mdm alone, so as not to limit our analysis capabilities for this study. It is likely that the additional cost of the a priori ensemble simulation could be reduced.

The paper summarizes results from many experiments, and therefore often requires re-reading of previous sections to capture what exactly has been done. The understanding would increase if the authors could include a table with all experiments and how their configuration is different, for example using the keywords R_a to R_e as used in Figure 11. In addition, a figure illustrating the data flow would be useful, as there are many entities used that could be confused with each other. There is for example the Kalman Smoother ensemble but also a meteorological ensemble; there are synthetic observations used but also real data; boundary conditions originate from ERA5, ICON, and CAMS; tracers could be either CH4 but also SF6; etc. A better overview of the data stream would help to distinguish the various elements from each other, and also help interested readers to implement a similar system themselves.

We have summarized all the inversions presented in this study in two tables. In the text, we now refer to the inversion IDs used in these tables. In addition to the ID, the tables list the type of mdm (flux-dependent or constant), the emissions used, the a priori and true variance, the scaling factor alpha used to achieve an innovation chi-squared value of 1, and the error reductions. For inversions with hourly and/or real observations, this is noted as a comment. However, we decided not to include an additional graphic to illustrate the entire data flow. We couldn't come up with a clear and concise schema that accurately depicted all the entities. Nevertheless, we believe that the restructuring and revisions made to the text in response to the reviewers' comments and the inclusion of the table have resulted in a clearer and more understandable manuscript.

Specific comment

Line 29: "Errors in the second term, commonly referred to as model-data mismatch error (mdm),".

A more common description is (I think) that **R** is the “observation representation error".

The matrix **R** is indeed often described as the observation representation error. Other papers refer to it as the model-data mismatch or simply as the observation error. We prefer the term model-data-mismatch here because a large part of the error is neither due to observation errors nor a representation problem, but rather due to model errors. To avoid any confusion, we now introduce the term as follows: "Errors in the second term, commonly denoted as **R** and referred to as observation representation error or model-data mismatch error (mdm), include all processes...".

Line 68-69: The term "OSSE" is more often used for to assimilation experiments that should quantify the impact of observation instruments or networks that do not exist yet. Not sure if the experiments with synthetic data that are described here could be described as such, but this is of course only a detail.

It is true that the term "OSSE" is often used to describe assimilation experiments designed to quantify the impact of proposed observing instruments or networks. Although our goal was different, the approach was the same, i.e. defining a true state (emissions) generating synthetic pseudo-observations, and performing inversions with these observations. It has become quite common in the community to use the term "OSSE" for such experiments.

Lines 87-88: What perturbed the CH4 boundary conditions? This seems described later on in section 2.3. Bud did the global ensemble use the same emissions and only different meteorology?

The CH4 concentrations in the global ensemble are perturbed not only by the differences in meteorology, but also by perturbed emissions, as detailed in Sect. 2.3. To clarify this at the point you mention in the text, we have added the following sentence:

"… as well as by perturbed CH_4 boundary conditions from the same global ensemble simulation that provided the meteorological boundary conditions. *In this global ensemble, the CH_4 concentrations are perturbed as a result of perturbed meteorology and perturbed emissions.*"

General: The order in which various elements of the system are introduced is more top-down than bottom-up now. Maybe this could be changed? For example: 1. model description (first paragraph under section 2); 2. meteorological ensemble (subsection 2.3); 3. ensemble simulations (subsection 2.1).

We have adjusted the structure as suggested. We have moved sections 2.3 (ICON-ART ensemble simulations) and 2.4 (CTDAS inversion setup) before the description of the two experiments and revised the text accordingly. We believe this restructuring makes the text more understandable.

Lines 103-104: "To ensure a fair comparison, inversions were set up so that the innovation chi-squared value in each inversion was 1". What does this mean, that the parameters that define the *a priori* model uncertainty are tuned to ensure Chi^2=1? Which parameters are these? It seems explained later in Eq.(1).

Later in the text, we describe that we achieve a chi-squared value of 1 by scaling the mdm and explain the reasoning behind this. To make this clearer already at this point, we have revised the sentence as follows:

"~~To ensure a fair comparison, inversions were set up so that the innovation chi-squared…~~" -> "To ensure a fair comparison, the mdm was scaled in each inversion so that the innovation chi-squared…"

In Eq.(1) the ratio is between scalars, but text mentions that **R** is a "covariance". Shouldn't it be actually a vector-matrix-vector product:

(yo-H(xb))^T (HPbH+R)^{-1} (yo-H(xb))

Yes, we corrected that.

Figure 6: The caption could include part of the description of lines 265-266 to know what the different lines are.

We have added the following sentence to the figure caption:

"Mean temporal correlations at Cabauw (a and b) and Monte Cimone (c and d) for observations at 00 to 08 UTC (a and c) and 12 to 20 UTC (b and d) with observations in the next 36 hours. *The error correlations over time (x-axis) for each observation is indicated by one line.*"

Technical corrections

The manuscript is well written and easy to read, so only a few errata found.

Line 105 and more: usually "*a priori*" and "*a posteriori*" are written in *Italic* font.

The ACP English Guidelines and House Standards explicitly state that common Latin phrases such as "a priori" and "a posteriori" should not be italicized. Therefore, we have kept the current formatting in the manuscript.

Line 129: .. observation*s*, ..

Corrected.