Reply to Referee #2 of egusphere-2025-1464

Valentin Bruch

October 23, 2025

We thank the referee for the detailed reading, positive assessment and valuable comments. Following a suggestion by the other referee and after consolidation with the editor, the manuscript has been split into a Part 1 and Part 2. In this new structure, we address the questions raised by the referee below.

The structure of this document follows the report of the referee. For transparency and convenience, we added Sect. 3 below to explain a mistake we noticed in the initially submitted manuscript during the review process.

1 Comparison to "standard" R values

Referee comment: "My main point is that I would like to see better visualised how the R matrix as made in this paper improves the inversion. The calculation is quite complex, and covers a large part of the paper. Nevertheless, some arbitrary choices remain (like the 10 ppb standard, and the inflation factor). I now wonder if 'standard' R values (like in Steiner et al., 2024b (reference as in manuscript): 'we use a model-data mismatch of 2 ppb + 40 % of the anthropogenic signal') would give similar results. [...]"

This is a very interesting question, which is now discussed in detail in Section 4 of the new Part 1. One main outcome of Part 1 is that the ensemble-based R matrix leads to a lower random error in the inversion results than the described "standard" R. But for comparing the different results, we choose a different focus than what the referee proposed:

Referee comment: "[...] For this, the χ^2-innovation can be used. χ^2 is used to assess the uncertainty, relative to the model-observation mismatch (i.e. $((y - Hx)^2)/HPH^T$, where y are observations, Hx are transported (prior mean) fluxes and P is the prior covariance matrix of the fluxes). One would like the χ^2-innovation to be as close as possible to 1.0 (which means uncertainty and model error are balanced). It would help the paper to include a figure of the χ^2-innovation to show that the 'new' R is an improvement over e.g. Steiner et al., 2024b. A run with a fixed (diagonal) R can also be included in Figure 7."

We agree that comparing different methods for constructing the R matrix based on the χ^2 innovation can be very helpful. But as pointed out by the referee, our method contains many tuning parameters. A simple comparison of the χ^2 innovation, as suggested by the referee, would only highlight differences between two configurations of tuning parameters but provide little insight into the method itself. Since we are interested in the potential of the ensemble method, we need comparable configurations and a way of assessing the inversion quality.

In the submitted manuscript, we already considered two methods for constructing R (denoted prior R and posterior R). In Part 1 of the revised manuscript, we add the "diagonal R" as a simple approach as also proposed by the referee (now in Sect. 2.5.1). One main focus of Part 1 is the comparison of these three ways of constructing R (prior R, posterior R, and diagonal R). However,

we consider the χ^2 method as not suitable for comparing the prior R and posterior R inversion. In particular, these methods have a different bias, which is not detectable from the χ^2 innovation. We therefore focus on different ways of comparing the methods but also mention the χ^2 innovation (now in Sect. 2.6.5 of Part 1, see comment below).

We contrast the methods using synthetic experiments with a simulated transport uncertainty (Sect. 5.6 and Appendix F in the first version, now Sections 4.2–4.4 of Part 1). Using these synthetic experiments with a known truth, we can compare how well a method can estimate the synthetic truth from pseudo-observations and thereby assess the quality of the inversion. We find that the diagonal R approximation without any information from the ensemble leads to a significantly larger random error in the inversion results. These results are visualized in the new Figures 6 and 7 of Part 1, which replace Figures 10 and F2. To make the inversion results comparable, we choose the tuning parameters such that all methods have approximately equal posterior uncertainties.

Referee comment: "It would help the paper to include a figure of the χ^2-innovation to show that the 'new' R is an improvement over e.g. Steiner et al., 2024b."

As mentioned above, we discuss the χ^2 innovation in Sect. 2.6.5 of Part 1. Instead of a figure, we chose a table (Table 2 of Part 1) as the best format to summarize χ^2 for the different methods for real and synthetic observations:

Table 2. Median of χ^2/N_{dof} for different configurations. χ^2/N_{dof} for the prior R inversion also serves as an approximation for the posterior R inversion. Synthetic observations are generated using the ensemble simulation, assuming that the a priori fluxes and the CH_4 concentration on lateral boundaries are known exactly.

observations	far-field correction	χ^2/N_{dof}, diagonal R	χ^2/N_{dof}, prior R
real	yes	0.18	0.16
real	no	0.21	0.18
synthetic	yes	0.05	0.03
synthetic	no	0.06	0.03

We find that the simple diagonal R has a larger χ^2 than the prior R, while χ^2 remains well below 1.0 for all cases. This result is closely linked to the constraint that the posterior uncertainties shall be similar, on which we have based the tuning parameters for the diagonal R matrix. This result indicates that our ensemble-based approach would allow for tuning to smaller posterior uncertainties than the diagonal R, then leading to larger χ^2. Nevertheless, this comparison does not imply the superiority of any particular method.

Referee comment: "A run with a fixed (diagonal) R can also be included in Figure 7."
We thank the referee for this constructive comment. Instead of adding only the diagonal R run to Figure 7, we compare all three considered ways of constructing R in the newly added supplementary Figure A4 in Part 2:

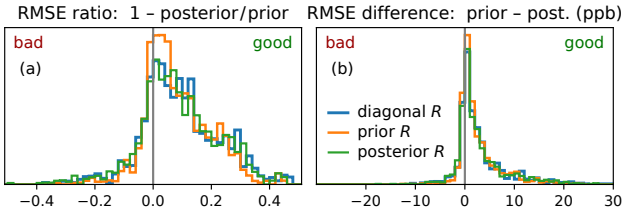

Figure A4. Statistics of the relative (a) and absolute (b) improvement of the model–observation mismatch at independent validation stations for different choices of the error

covariance matrix R discussed in Part 1. The figure is analogous to Fig. 5 [previously Fig. 7], where the visualization and the data selection is explained. Here, we distinguish three inversion methods that differ in how R is constructed, as introduced in Sect. 2.5 of Part 1. No clear advantage of one method over the others can be seen. The diagonal R inversion has the lowest posterior RMSE at validation sites, followed by the posterior R and prior R inversion, but the differences are not statistically significant.

In the new structure, suggested by the other referee, the discussion of the method is separated from the discussion of the main results. Since the validation (Figure 7) belongs to the results in Part 2, we have included this as a supplementary figure of Part 2 and not in the main text.

2 Other comments

1. *Referee comment: "Almost all IDs tested are within 15 % in their flux-solution. To me, this seems like the posterior uncertainty is under-estimated. Can the authors comment on this?"*

   The sensitivity tests typically show variations within 15 % of the posterior 2σ uncertainty. This indicates that the uncertainty due to the choice of tuning parameters is small compared to the posterior uncertainty. Based on this finding, we do not try to include the uncertainty due to the tuning parameters when estimating the posterior uncertainties. We see no evidence that the posterior uncertainty might be considerably underestimated.

2. *Referee comment: "Can the authors explain to me the strange boundaries in Fig. 4b and d? With this, I mean the darker lines that run through e.g. France."*

   When defining flux categories by area, we used a smooth transition at boundaries between different categories. For example, we split Poland into two flux categories (west and east), but some emissions in the center are assigned to 50 % to western Poland and 50 % to eastern Poland. This setup helps us reduce sharp spatial concentration gradients in our transport simulation. However, when assuming uncorrelated a priori uncertainties for eastern and western Poland, this implies that emissions that belong to both categories have a lower relative uncertainty. These boundaries between flux categories are therefore visible in the uncertainties. We added a brief explanation in the figure caption: "The smooth boundaries between two regions with separate scaling factors appear as darker lines because these scaling factors are assumed to be initially uncorrelated."

3. *Referee comment: "in Section 5.5.1, a gaussian noise of 2 ppb random error is added to the pseudo-observations. However, the σ_{const} is already 10 ppb, which means the added white noise is quite small compared to the uncertainty associated to these observations. Can the authors explain this choice?"*

   In the synthetic experiments with randomized true emissions, we work with idealized pseudo-observations that have an underestimated error. The simulation of a transport error for these pseudo-observations would require an impractically high computational effort. Thus, these pseudo-observations inevitably only include a very simplified error. But this merely impacts the analysis, since the focus of these experiments lies on testing how the full inversion system – including observation filtering and far-field correction – can determine the synthetic truth from idealized pseudo-observations.

   In lack of a good estimate for the error on the pseudo-observations, we choose the values 2 ppb which is larger than the observation uncertainty (of usually < 1 ppb) but sufficiently small to

not impact the outcome of the synthetic experiments. To clarify this aspect, we have extended the explanation in Sect. 5.5.1 (now Sect. 3.4.1 of Part 2):

> This construction of pseudo-observations clearly underestimates the true error in the model–observation comparison, but it allows us to test the interplay of far-field correction and inversion in a controlled setup. Synthetic experiments with a simulated transport uncertainty are discussed in Part 1.

In this work, we consider different types of synthetic experiments which either consider a simulated transport error or a random variation of the emissions. Combining both would be an interesting extension, but goes beyond the focus of the current study.

3 Mistake in the manuscript

Figures 10, F2 and E1 in the submitted manuscript were affected by a mistake: Those inversions for which the far-field correction was disabled used wrong tuning parameters with underestimated uncertainties due to a bug in the code. In sensitivity test 400 and Figure 10, this led to an overestimation of the relevance of the far-field correction. The mistake has been corrected and the discussion of the far-field correction has been adjusted accordingly. The conclusions of our manuscript remain unaffected.

(This paragraph is included in the replies to both referees.)