Reply to Referee #2

Referee comment in bold, reply in plain text, *modified text blocks for manuscript in italics*.

Please note that the line number given by the referee refer to the originally submitted manuscript and not the one published on ACPD.

The manuscript "Validation of the Swiss methane emission inventory by atmospheric observations and inverse modelling" by Henne et al. analyze an ensemble of atmospheric inversions of the Swiss methane emissions using continuous CH4 measurements from a ground based network of 4 to 6 sites. The study highlights the consistency between the results from the different inversion set-up and the SGHGI inventory. Finally, it attempts at interpreting the spatial and temporal distribution of the inverted emissions, linking it to the different types of sources and processes underlying the CH4 emissions in Switzerland.

In general the manuscript is very well structured and written. The inversion system and the experimental framework seem robust, and the results are highly encouraging for the use of atmospheric inversions as a mean to verify and potentially improve the GHG emission inventories. I do not see any major issue in this manuscript, which I strongly recommend for publication in ACP.

I still have some concerns regarding the text and the results as detailed below.

We would like to thank the reviewer for this very careful review that clearly helped to improve our manuscript.

There is a series of minor issues related to the "baseline treatment". The definition of the "baseline" and of what is targeted through the optimization of the "baseline" is unclear. The term baseline is not really self-explanatory. The definition of the baseline as "conditions without recent emission input" (l. 186) is vague even though it connects it to the concept of model boundary conditions. The different configurations of the state vector for the baseline are not really justified by a characterization of this baseline. Why could it be assumed to vary between the sites that are often quite close to each other compared to their distances to the boundaries of the inversion domain? Why could it have a full 3D structure in space? The impact of including the baseline in the state vector and the sensitivity to its configuration is important for such a regional configuration. When the inversion is allowed to optimize a different baseline for each site, it can easily make attribution errors between the baseline and the emissions, as illustrated by the results from experiments S-ML (there is a beginning of discussion on this topic at lines 459-462). Most of the inversion tests use low prior uncertainties in the baselines, which could limit the influence of the baseline on the inverted emissions, but which, on the other hand, may prevent from fully accounting for the uncertainties in the boundary conditions of the inversion domain. The baseline should thus be better defined and discussed, its weight on the inversion and the corresponding attribution errors should be better characterized.

The baseline concentration as seen by our regional scale LPDM can be understood as the average mole fraction at the individual particle locations at the end of the transport simulations (4 days before arrival at the sites or at a domain boundary). As such this baseline mole fraction is not simply the mole fraction at the domain boundary but may well contain contributions from within the domain in cases where our backward integration time was not sufficiently long for all particles to leave the domain. The justification for using an individual baseline for each site lies within this fact. In our most simple baseline treatment (used by various authors before) any information on the particle end points is neglected. Even though the baselines for nearby sites should be similar they may still vary considerably depending on the elevation of the site. This is especially true for the sites with very different elevation JFJ, SSL versus BEO and LAE. The 3D structure of the baseline presents the attempt to deduce the mole fraction at the particle end points as one common baseline.

The potential for attribution errors was the reason why we introduced 3 different treatments of the baseline. Since these yielded similar results we had hoped to refute any concerns regarding this point. However, we see that it merits further discussion in the manuscript.

We extended the description of our baseline definition and highlighted the different considerations, given above, in the revised manuscript (sections 2.2, 2.3, 2.5.7).

We also point out the potential of attribution errors due to the baseline treatment. We discuss the results of two additional test inversions with factor of 2 lower or higher baseline prior uncertainty. Their result for total Swiss emissions was 210 Gg/yr and 160 Gg/yr for low and high baseline uncertainty, respectively, and low particle release height. This is within the range defined by all other sensitivity inversions. Given the fact that in neither case the obtained posterior baselines looked very realistic (too smooth in the case of small prior uncertainty and too closely following the observed variability in the case of large prior uncertainties), we did not include these cases as part of our sensitivity inversions but discuss them along with the alternative baseline inversions.

I feel that the authors go too fast in rejecting the assumptions that the positive corrections applied by the inversion in the north-east of Switzerland could be artifacts. The fact that those corrections could balance (through attribution errors or due to a lack of flexibility) the imperfect inversion of the baseline for north-east winds is quite ignored while this is a likely explanation. Furthermore, I do not agree with the sentences at line 908 "it seems unlikely that the same systematic bias would be inherent to both meteorological inputs" and at line 22 "which rules out an artefact of the transport model and the inversion system". Both meteorological forcings have a limited spatial resolution (which can be critical in this area) and errors inherent to FLEXPART would apply to both FLEXPART-ECMWF and FLEXPART-COSMO. Section 4.3 fails to identify a significant missing source. Therefore, my feeling is that the discussion attempting at linking the corrections in the north-east to specific processes in section 4.3 is a bit too long and a bit too ambitious. It could be shortened and the text should acknowledge that some limitations of the inversion system could explain such corrections. What is the weight of these corrections in the national budget of the emissions from the inversions? Considering that this is connected to a sector underestimated by the SGHGI, or to an artifact of the inversion, considering that the inversion decrease the emissions from agricultural lands compared to the SGHGI, and given that SGHGI is a purely bottom-up estimate, it seems that the fit between the national budgets from the inversion and from SGHGI could be seen as a coincidence without such a weigthing.

We understand the concerns of the referee. We hope that some of them were already addressed by the more detailed explanation of the baseline treatment and its uncertainties (see above). We followed the advice of the referee and don't rule out inversion artefacts in the manuscript anymore. However, we did not shorten the discussion on the potential source, especially since also referee 1 highlighted the potential role of different manure handling practices that could induce large spatial differences in this emission type.

The additional test inversions with factor of 2 different baseline uncertainty did not change the spatial pattern of emission changes in Switzerland significantly. The increase in north-eastern Switzerland was confirmed also for the inversion with large baseline uncertainty that resulted in considerably lower (160 Gg/yr) country total emissions than the base inversion (similar to results from S-ML).

One additional argument in favor of a true increase in north-eastern Switzerland is that it is also observed in summer time when pollution events with easterly advection are absent (see Figure 6).

The contribution of the gird cells in the north-eastern part of Switzerland (selected 4 grid cells with largest increase in posterior) to the national total emissions to total Swiss emissions was 16.3 Gg/yr or 9 %. In the posterior this was increased to 22.5 Gg/yr or 12.5 %. The change of 6.2 Gg/yr represents 3.4 % of the national total. This is not a negligible contribution to the national estimate, but does not compromise our conclusion that the posterior estimate is in close agreement with the bottom-up inventory, even if the increase in the north-eastern area is ignored completely.

Several components of the inversion configuration need to be presented in a more explicit way and earlier than they are presently: the temporal resolution at which the emissions are estimated (it should be explained in section 2.3, and anticipated at the end of section 2.2 at lines 246-. . .), the transport modeling and inversion spatial domain (what are the COSMO-7 domain and Western Europe in section 2.2 ?). Section 2.3 should confirm the inversion period (Mar 2013-Feb 2014) even though it is implicitly defined near the end of section 2.1. Much of the model vs. data analysis are lead without stating whether they are based on 3-hourly timeseries with or without the data selection applied to the data assimilation (using 12:00-18:00 and 0:00-6:00 time windows only, and filtering the data according to the wind speed: line 77 states that it is used for the inversion, but we do not know whether it is used e.g. for the REBS analysis few lines later). Typically, the text does not say precisely what is plotted in figure 3 and 4 in terms of temporal resolution and temporal selection.

We followed the advice of the referee:
We included the information of the temporal resolution of the target emission estimate.
We added information on the meteorological domains.
We confirmed the target period in 2.3.
We clarified that only the 3-hourly filtered and aggregated time series were used in all inversions, analysis and discussion.
We added missing information on temporal resolution to Figure 3 and 4.

Minor comments

line 5 vs line 9 vs line 20: you should introduce the "reference" prior emissions at line 5
The reference prior emissions are now introduced first.

line 9 and elsewhere: use error (or uncertainty) covariance instead of covariance by itself; use something more self-explanatory than "baseline treatment" at line 10 ?

We use uncertainty covariance throughout the text now.

We replaced the term baseline in the abstract by “large-scale background mole fraction”. Later in the text baseline is explained in detail and we stick to this terminology.

line 8 vs 12 vs 15: you should use the same designation for the inventory, otherwise the reader can believe that you speak about 3 different inventories

We stick to SGHGI throughout the whole text now.

- line 11: I feel that "independent character" is a bit strong, even though I do not contest the robustness of the system. You still have large sensitivities to e.g. the prior uncertainty in the baseline which is a critical component of the system. Furthermore, here, you do not necessarily speak about the national budgets only, and the sensitivity of the spatial distribution of the emissions to the inversion set-up is also significant.

Although we used relatively small uncertainties for the prior, we established that different prior distributions and totals (MAIOLICA vs. EDGAR) did not result in significantly different posterior emissions. Also increasing the prior uncertainty as done in S-ML did not result in significantly different posterior emissions. The same is true for the baseline. We investigated various baseline treatments in order to explore the range of possible baseline influences. Considering these results from the sensitivity inversions, we are confident about the independent character of the inverse estimate, which showed only small effects due to assumptions on the prior.

- line 58-59: "Methane emissions from individual sources are much more difficult to quantify than anthropogenic emissions of CO2" it's a too general statement to be true

We agree that for individual sources this is not generally true. We were rather referring to the fact that total national anthropogenic CO2 emissions can be derived relatively precisely because combustion is by far the most important anthropogenic CO2 source and fuel statistics, for a country like Switzerland, are very well know . In contrast, CH4 emissions stem from less well understood processes that may vary strongly on an individual basis (ruminants, manure handling, waste treatment). We revised the wording of this sentence to clarify this point.

- l. 71-72 : this raises some questions; inventories of CH4 emissions should be built on such an upscaling of site / process scale measurements or on highly uncertain emissions factors, so why would this be too difficult ? or why were the numbers from these specific studies difficult to upscale (because the sampling was too small ?) ?

These studies did not necessarily focus on an individual emission sources but rather on small regions. Therefore, they did not provide emission factors for a source process that could be up-scaled. Due to different source compositions within each analysed area and the small number of studies it is not easily possible to translate these results to the country scale. However, we revised the wording, removing the word up-scaling, to clarify this.

- line 75: the inversion delivers but does not "combine" an optimal estimate of the emissions
We revised the sentence accordingly.

- line 92 : evaluate instead of validate ?

We would like to stick to validate in this context (see also title of manuscript). The discussion about validation versus evaluation was raised by the modeling community due to the fact that a model can never be "validated" (it will always be wrong to some extent"), but it is possible to evaluate whether a model is fit for a given purpose. However, we are validating the inventory in the sense that we are checking whether the reported emission total is consistent with our (more or less) independent estimate *within the range of the combined uncertainty*.

- l. 95 give an idea about this high resolution
Information was added to the manuscript

- l.114: you do not target the Swiss Plateau in the following, but rather the national budgets which is thus something different (by 30%)

The sentence states that the sites are "mainly" sensitive to the Swiss Plateau, but not exclusively. Furthermore, we add two more sites in the inversion (JFJ and SSL) which are sensitive to larger areas than our Swiss Plateaus sites. Overall the combined sites should be sufficiently sensitive to the whole of Switzerland to derive national emissions. In addition, the inversion will account for the lack of sensitivity (for example south of the Alps) by not reducing uncertainties in these areas.

- l. 185, the justification for using nighttime data at JFJ and SSL goes too fast and should explain why the situation regarding the PBL and the local sources is different there compared to the other sites (otherwise it could seem in contradiction with lines 182-183).

We extended the explanation why we used night-time data from JFJ and SSL.

- l 185-198: this paragraph is not fully clear (is the baseline an annual mean, or a series of 60 day averages ?) and will generate some confusions in the following, where the baseline will rather be computed as an interpolation of "baseline nodes" and where only the REBS analysis for JFJ will be used. You should find a way to avoid this little issue either here or later when describing the inversion protocol. "reflecting different degrees of variability and frequency of air masses not influenced by recent surface contact and emissions." is not really clear.

We added some clarification, but refer to the REBS publication for further details on the method. We furthermore added a sentence on how the mole fractions at baseline nodes were derived from the REBS estimates in section 2.3.

- l. 206: why the spatial resolution of ECMWF is degraded outside the Alpine area ?

for computational issues ?

This is due to limited storage capacity and done in a similar way by many groups using similar modelling approaches). Away from the releases points and the steep topography the coarse ECMWF resolution still gives sufficiently accurate transport results.

l. 208-210 : can you explain the role (on the top of the analysis from MeteoSwiss) and horizontal resolution of COSMO ?

COSMO is the NWP system used to generate hourly meteorological fields using data assimilation of observed parameters. We did not run COSMO ourselves, but merely used the operational product of MeteoSwiss. The main advantage of this, as compared to doing independent COSMO simulations, is the benefit of the assimilation cycle and the continued validation efforts by MeteoSwiss in terms of meteorological variables. We use COSMO analysis at a 7 km by 7 km horizontal resolution to drive our transport simulations. Although this resolution is not entirely sufficient to describe all details of the complex, near-surface flow in mountainous terrain, it proved to be reliable for the elevated sites used in this study. This information is given in some detail in section 2.2, but we added additional COMOS domain information to the revised text.

- l. 214 "time-inverted" was not mentioned before

We now mention this when the FLEXPART versions are introduced.

- Table 1: the particle release heights are given for FLEXPART-COSMO only, not for FLEXPART-ECMWF unless the release heights at JFJ and SSL for the latter are based on the ground height in COSMO ? (the paragraph on the release heights is not fully clear about this).

We added this information to the table.

- l .244: for the mountaintop site JFJ during nighttime only ?

The cited publication evaluated simulations for all times of day and night-time separately and did not establish significant differences due to this selection.

- l. 258-259 : you speak about sensitivities, not about sensitivities multiplied by the emissions, so you should not write lines 257-259 (or you should speak about mi,j times Ei,j and defining Ei,j correctly)

We reformulated this sentence.

- l. 279-280 should be improved

We understand that the referee requests further details on how the grid box aggregation was done. We added these to the text.

- l. 284: can you give more details about the domains that were tested? it could be important for the discussion on the baseline and on the corrections in the North East of the Alpine area.

The current inversion domain has an eastward extent up to 21° East, this is also roughly the extent of the COSMO7 domain, which was used for the FLEXPART simulations. Reducing this eastward extent by 5 degrees did not show significant differences in the inversion (both in terms of baseline and emission adjustment). With a further reduction in the eastward extent, the under-prediction during the highlighted pollution events increased and emission increments towards the east were more emphasised. We did not include these experiments in the paper in more detail since the choice of a sufficiently large inversion domain seemed to be a prerequisite for a successful inversion. The CSOMO7 domain limits are now given in the text.

- l. 327 : you will correct the baseline differently for each site, so why using the REBS analysis at JFJ site as a prior for all sites ? see the main comment regarding the lack of characterization for the baseline which could help justify the way it is accounted for in the inversion state vector and the way priors and prior uncertainties are assigned to it.

We used the baseline from JFJ since this was most robustly obtained from the observations. The REBS method generally requires that about 50 % of the observations are baseline observations in the sense that they follow a normal distribution about a smoothed baseline curve. For the sites closer to emissions this condition is usually not met. While REBS still produces a reasonable baseline curve here, it is evident that it often lies above the lowest observed mole fractions: a clear indication that the method did not fully succeed. We did a test inversion using the REBS baselines obtained from each individual site that did not succeed in terms of bringing down the posterior baseline, resulting in simulated mole fractions that were generally too large during little polluted conditions.

- l. 303: "the vector f gives the fractional contribution of the region to each inversion grid cell" something is inverted in this sentence and it could be said better in order to mention that it is related to the surface area (see l. 303 vs. l. 306-307). l. 306-307: I am not sure to understand, you will thus derive xk in terms of emission per capita instead of per m2 to be consistent ? then will it impact the results ? The whole paragraph should be clarified.

The calculation of f_k in grid cells that span more than one region is based on high resolution population density and not simply on a by-area contribution. This is to avoid wrong allocation in cases where for example a large city of region A is situated in the same inversion grid cell as a sparsely populated area of region B. In the case of the current CH4 inversion the influence of this treatment is of minor importance. It was introduced in our system for application to halocarbon inversions.

We clarified the description in the manuscript.

- l. 317: the letter f is already used for eq. (7), you should use another letter to avoid confusion ?

Here it is f_E not f. However, we changed the letter f in equation 7 to g to avoid confusion.

- Equation 8 and associated parameters at lines 322-323: you do not discuss and justify this modeling of the spatial correlations. Ignoring the independence between the different types of sources of methane in such a definition could raise problems. The 14-day timescale for the errors on the baselines, which could miss the signal associated with synoptic events, is also given without justification. The set-up of the prior and observation/model error covariance parameters cannot be perfect but it could be supported by few explanations.

We are aware of the fact that this approach is a simplification, but it has been widely used in lack of more detailed information on the spatial correlation and in order to further constrain the posterior solution (e.g., Rödenbeck et al., 2003;Gerbig et al., 2006;Thompson and Stohl, 2014). Hiller et al. (2014) investigated the spatial correlation length scale by analysing differences between the MAIOLICA, EDGAR and TNO inventories in a variogram approach. Their estimate of 13 km and 8 km when comparing to the EDGAR and TNO inventory is considerably lower than what we used in the inversion. When using a value around 10 in our inversion, emissions in individual grid cells become almost independent of each other and the prior did not provide sufficient constraint for a reasonable posterior solution, as indicated by seemingly arbitrary posterior adjustments and the introduction of dipole structures in the posterior field.

The chosen length scale for the baseline represents a compromise between something that would be closer to the synoptic time scale (~5 days) and the need for additional constraint on the baseline. Values lower than 14 days tend to give too much freedom to the baseline. In addition, we would argue that most of the synoptic-scale variability should be picked up by the advection changes within the domain (intensity and location of contact with the surface/emissions) and not the baseline, given the rather large extent of the inversion domain.

These considerations are not free of a certain degree of subjectivity. However, both parameters were evaluated in the ML method later in the manuscript and similar values were obtained, lending credibility to our choices.

We give additional motivation for the use of the chosen parameters in the manuscript.

- l. 321: if you use the vector 1, you will include emissions from all countries, while you aim at comparing it to the Swiss budget (line 323). Could you confirm that you scale uncertainties outside Switzerland based on the scaling factor derived in Switzerland?

Yes the same relative uncertainty is applied for grid cells outside Switzerland as can also be seen in figure S2.

- l. 330 add "at a given site"; and speak about the spatial correlations between errors on baselines at the different sites (some justifications may be needed). Line 333: then you apply the uncertainty diagnosed by the REBS for the annual baseline to the uncertainty in 5-day baseline nodes ? or does the number given by the REBS mean something else ?

We added the suggestion to the text and keep the assumption of no correlation between sites (see argumentation in reply to first question.

The uncertainty provided by REBS does not give an annual mean but the prediction uncertainty at every time step so that it should be appropriate to apply it to the 5-day baseline nodes.

- the paragraph regarding the model errors is quite confusing and should be improved; it is difficult to understand what exactly means "a constant contribution while the third term represents an uncertainty contribution relative to the prior simulation of above baseline concentrations". Through these computation aren't you attributing part of the prior uncertainty to R ? Lines 342-348 go too fast and are unclear. "Residual" is not a self-explicit term.

We improved the description of the concept. Generally, we agree with the reviewer that in our method the prior has some influence on the estimate of R. We added a note of caution to this extent:

These methods have in common that the results of the prior simulation influence the estimation of ***R*** *therefore somewhat violating the independence of prior and model/observation uncertainties assumed in the Bayesian approach.*

However, the choice of the method is a compromise driven by the lack of better information to characterise the model/observation uncertainty.

-l. 356: the selection of the data should remove this problem of the simulation of the diurnal cycle in the PBL ? furthermore, such an error sounds like a "bias" while 0.5 days is relatively short

We agree with the referee that by selecting only day-time (night-time) data the problem of the diurnal cycle is diminished in our inversion. However, we use two afternoon (or in case of SSL and JFJ nighttime) observations per site and day and therefore the correlation length scale is still appropriate to take care of auto-correlated nature of the observation errors. Others have also determined the temporal correlation length scale from the auto-correlation function of the prior residuals. When we attempted this, values close to 0.5 and no significant changes on the inversion result were obtained.

- l. 362-363 : even if these assumptions were true, the set-up of the inversion would still be highly uncertain, which explains most of the sensitivity tests. What do you mean by "uncorrelated residuals" ?

To explain what we mean by “residuals” we changed line 342 to “were estimated separately for each site from the model residuals (differences between simulated and observed values) of the prior simulation” and changed “uncorrelated residuals” on line 362 to “temporally uncorrelated residuals”. Most of our sensitivity tests are addressing potential systematic biases. If the assumptions listed on lines 362-363 were true, most of these tests would not be necessary.

- l. 385: could you define the seasons (the corresponding months) here ?

We included a definition in the text.

- l. 385 the text could mention here or even before (when defining the temporal resolution of the state vector) that the emissions may actually have significant variations in time (even though it will be discussed in section 4)

We included a statement at this location.

- l. 389: here as in section 2.5, you describe some new theoretical components faster than in section 2.3 while it could require the same level of detail (especially when presenting the extKF). I suggest to (1) generalize section 2.3, (2) detail the configuration for the base inversion (3) present the sensitivity tests by detailing each corresponding perturbation to this base inversion using the mathematical framework of 2.3.

We followed this suggestion and extended 2.3 at the expense of somewhat shortening 2.5. However, we tried avoiding duplication of information wherever possible.

-l. 390: 90 days is nearly one season, so the inversion might have a limited ability to increase the seasonal variations. Do you know whether decreasing this temporal correlation would have significantly increased the seasonal variations that are analyzed in 3.2 and discussed in section 4 ?

When reducing the temporal correlation length to 45 days changes in the seasonal posterior emissions were not more than +/- 4 Gg/yr compared to the setup with a correlation length of 90 days. We consider this a minor influence, but added this information to the text.

- Section 2.5.3 is a bit short given that many of the inversion parameters, concepts and underlying assumptions change when switching from the base method to extKF. And the text should explain which type of uncertainties are targeted by this sensitivity test. This section should better characterize what is the state vector for each (3-hourly ?) analysis of the sequential algorithm. What is the time resolution of the corrections to the emissions here ? L. 396: what is the tendency of the baseline value ? Are there some consistencies between the B and R matrices used here and that used in the base inversion ? Do you need the ML method to set-up these parameters (maybe switch the sections 2.5.4 and 2.5.3)? l. 403-405 is impossible to understand; how Q is set up ?

We kept the description of the Kalman Filter short on purpose, since it was recently published in full detail. This is in contrast to the implementation of our Bayesian approach, which we never published in as much detail as presented here. However, we added some additional information on the Kalman Filter approach to improve the understanding of this section, in particular regarding the design of the uncertainty covariance matrices **B**, **R** and **Q**.

- L. 408: based on the national emission uncertainty as estimated by SGHGI provides the best knowledge of the uncertainty in SGHGI, not of the uncertainty in the prior estimates (MAIOLICA and EDGAR) used here

We clarified this in the text. It now reads:
Our base inversion is based on the prior emission uncertainty as estimated by the SGHGI, which we consider to be the best knowledge of bottom-up uncertainty in Switzerland. Since Hiller et al. (2014) used the same by-category emissions as the SGHGI to spatially disaggregate total emissions for the MAIOLICA inventory (our prior), we extrapolated the SGHGI uncertainty information to the whole inversion domain.

- L. 415-416 : residuals and differences sound like shortcuts to me

We are following standard terminology: The term "residuals" is used strictly for differences between simulated and observed mole fractions. This is now explained in the revised text at the first place where "residuals" appears.

- L. 427: why ? in theory, the model/obs errors should be similar in the base and extKF inversions, even if the changes in the state vector can impact part of the model error (e.g. the aggregation errors related to the resolution of the state vector)

This is merely due to technical reasons. It simply was not implemented in the system yet. We added this information to the text.

- l. 435 : who reports to UNFCCC ? FOEN and thus MAIOLICA would be consistent with SGHGI ?

FOEN is the Swiss Office for the Environment which annually reports the SGHGI. The reference for the SGHGI is thus FOEN (2015). The total emissions in the MAIOLICA inventory for the anthropogenic sectors are identical to the SGHGI (reported in 2012). Natural emissions were added in MAIOLICA but estimated to be almost negligible. The total uncertainty of the SGHGI thus also applies to MAIOLICA. We clarified this in the text.

- l. 445-446: rewrite, it's hard to understand.

Indeed this was not well formulated. Actually we were referring to the total uncertainty of the country emissions. The revised text should make this clear now.

- Sections 2.5.5 and 2.5.6 define some critical components of the base inversion. So, in section 2.5, we are not just looking at perturbations to the base configuration. This supports my earlier suggestion to start giving a general framework for the inversion which could apply to both the base and sensitivity cases, and then to present the base and sensitivity configurations.

The information on prior emissions in the base inversion was moved to section 2.3.

The treatment of the baseline is now more generally introduced in section 2.3. However, the details of the alternative approaches are still detailed in section 2.5.6.

- L. 465 and 469: "initial" may be confusing here

Changed to *FLEPXART particle end points*.

- L. 463-465: I would rather discuss the link between the baseline and the boundary conditions (which can be derived from large scale transport models) earlier when defining the baseline.

We do this in the revised manuscript at the end of section 2.2

- L. 473 to 485: shouldn't the baseline rather be discretized in space over the inversion domain boundaries ? The concept of a 3D discretization of the baseline looks strange (see the general discussion about the baseline), but it's balanced in this inversion test by the fact that the same baseline is applied to all sites, and by the coarse resolution of the grid for this 3D baseline.

See general discussion above.

- L. 481-482: it is surprising to see that this level is at 3000magl. What is the typical PBL height during the afternoon in winter / summer ? One could have thought that most of the sensitivity for afternoon data lies in the PBL. Of note is that it could make sense to select a vertical separation close the PBLH.

The typical diurnal maximum boundary layer height is below this altitude (~1500. m agl in summer). We first used this as a threshold, but observed much larger sensitivities in the sub-column above 1500 m than for the sub-column below. This can be explained by the fact that we are looking at the particle end points (4 days before arrival at the sites) where particles were already more evenly mixed throughout the troposphere. The low sensitivities in the below 1500 m sub-column did then result in minor adjustments of the prior values, which seemed unreasonable. Hence, we chose a threshold that distributed the sensitivities more evenly in the vertical.

- l. 504-505: you do not explain this mapping so this is not straightforward

This mapping artefact is due to inversion grid cells covering more than one country and therefore emissions may be assigned to the wrong country. We tried to remedy this by using a by-population weighting in such grid cells (see discussion above), but could not get rid of this effect completely. Further steps could include an inversion grid design that is oriented along country borders or a higher resolution inversion grid. We extended the sentence to explain that the mapping artefacts are associated with grid cells along the Swiss border overlapping the neighbouring countries.

- l. 517, 519: this seems more frequent than induced by these sentences. Oct 2013 is an example comparable to March April 2013.

We added the Oct 2013 period to the text.

- L. 525-526 : we do not know (in the text; the figure gives the answer) if this prior baseline is a REBS analysis of the prior simulation, or the prior of the baseline state variable which should be equal, for all sites, to the REBS analysis at JFJ

We added this information also to the text.

- L. 527 and 543: Figure 4 does not necessary indicate a considerable improvement of the fit to the data after inversion. However, the inversion relies on corrections to the emissions and baseline at very low time resolution, so it has a limited potential for increasing the fit at high temporal resolution, which could explain why the posterior correlations hardly exceed 0.6. So would not it be more relevant to check the correlations between the model and the observations at a lower temporal resolution ?

Note that values of R^2 are shown in Figure 4, not of R. The posterior correlations at the sites BEO, LAE, SSL and JFJ used in the inversion reach values between 0.75 and 0.83 and not only of about 0.6. The reason for limited improvements can be seen in the good quality of the prior inversion both in terms of emission strength and spatial distribution. The improvement is still visible even if one does not want to call it considerable. However, we would argue that these improvements are considerable, since they explain an additional 15 % of the observed variability as compared to the prior simulation.

We already discuss correlations for above-baseline (high frequency) and complete signal and indicate that some of the improvement of the complete signal originates from the baseline adjustment. The improvements in high frequency can still be significant in a grid cell inversion since a spatial redistribution of emissions can help improve also the simulated temporal variability.

Furthermore, the temporal correlation in the residuals is only about 0.5 days suggesting that we can extract independent information from the observations up to that frequency. It might thus make sense to reduce the temporal resolution of the comparison to 12-hourly means but not to anything longer. This would further improve the correlations somewhat but we prefer to stick to the full time resolution in the manuscript.

- l. 533: you have a major part of North Western and Central Europe in your domain

While we have large parts of North-Western and Central Europe in our domain, we did not claim here that the flow was from these regions. We said that the high pressure centre is over Central Europe that means that the flow towards Switzerland was from Eastern Europe and possibly originated outside our model domain.

- l 536: I don't really see it at JFJ

The statement was supposed to refer to the fact again that the baseline was derived for JFJ, but we agree that it is confusing and reworded it.

- l. 562-563: I'm not sure to understand what indicates the link between the correlation and "general flow to the site", and the link between the STD and the local processes

We assume that the temporal pattern of the variability is driven by transport, whereas the amplitude is driven by the emission strength and initial vertical mixing (in turn the PBL) height. The model captures the temporal variability well (correlation) but does not capture the amplitude (std), hence our suggestion. We clarified this in the revised text.

- l. 585: this study uses a higher resolution modeling so the scores should be better; but it also depends on the type of sites where these scores are derived (this study focuses on a complex area)

As the referee correctly states, we use high resolution transport modelling compared to the cited studies, but include sites in complex terrain close to emission sources, which means that these will show more complex variability than those used previously. The fact that our model system is still able to simulate the observations as well as more remote sites in other studies, documents the validity of our system. We emphasised this fact in the revised manuscript.

- l. 603: - 605: putting these 2 sentences together is quite confusing. In one hand you discuss the theoretical uncertainty reductions. On the other hand you have an estimate based on sensitivity studies exploring sources of errors ignored by the theoretical computation (here the uncertainty in the particle release height). Mixing these two numbers (you say "additional uncertainty range") is a bit confusing.

We understand the second sentence as a note of caution at this point that the purely analytical uncertainty should not be taken as the only uncertainty of the posterior estimate. We clarified this in the revised text.

- L. 640, 665, 697. . .. the maps of posterior emissions are systematically stated to be "similar" to that of the base inversion. It is sometimes difficult to assess the level of similarity since the supplementary material does not show these maps (it already shows a lot of material). But the maps of corrections to the prior emissions by the inversion sometimes strongly differ regionally (such as S5c), S8c) S17c) vs. 2c)). The uncertainties in the inverted emissions for some specific

areas may thus need to be more emphasized in the discussion sections when analyzing the spatial distribution of the emissions.

Since for most sensitivity inversion the same prior was used as for the base inversion it should be sufficient to look that the corrections (posterior differences).

For the S-K sensitivity inversion (Figure S5) we outline the similarities and differences to the base inversion in lines 642 to 645. For S-E (Figure S8) a different prior is used as explained in the text. Therefore, it is not enough to just compare the corrections in this case. However, the posterior distribution looks very similar to the base inversion as shown below. As the referee stated, there is already a lot of information in the supplement. Therefore, we did not want to include an additional figure at this point.

For inversion S-B2 we included an additional discussion of the differences in posterior distribution in the text.

- l. 647-648: as previously discussed, the low temporal scale of the state vector in the base inversion limits the potential for increasing the correlation to 3-hourly data. extKF correcting the emissions at a much (which one ?) higher temporal resolution, it definitely has a higher potential for increasing this correlation.

The Kalman filter corrects the emissions and baseline on a daily basis. We already state the fact that this is the cause of increased posterior model performance in the text but put more emphasis on this in the revised text. In the revised version of section 2.5.3 we describe the setup of the matrix Q (prediction uncertainty) in more detail and mention that the values chosen in Q only allow for very slow temporal changes in the emissions that do not even fully capture seasonal fluctuations. The higher skill scores are thus only to a small extent due to the temporal variations of the emissions but are mainly due to using 6-hourly averaged (instead of 3-hourly) values and a more flexible treatment of the baseline.

-l. 657-658: clarify and explain it earlier when presenting the extKF in section 2.5.3

We added this information also to section 2.5.3.

- l. 678: this is not straightforward; the theoretical computation of A can account for the model performance only through the set-up of R. Then this discussion should rather be based on the comparisons of the estimates of R for COSMO vs. ECMWF.

On second thought we agree with this statement and removed the sentence and replaced it with: *This can partly be attributed to the larger model uncertainty assigned in the ECMWF case (especially low particle release case) compared to the base inversion (compare Tab. 3).*

- l. 681-684 the link between smaller posterior emissions and the diffusivity is not fully straightforward and could be better explicited.

The explanation is given in the following sentence. Increased vertical and horizontal dispersion of an emission plume will lead to lower atmospheric mole fractions when the plume intercepts a receptor. If the model under-estimates the dispersion the plume will be more concentrated in the simulation as compared to the observations. In order to align model and observation a decrease in emissions would be necessary.

- l. 705-708: the high corrections and uncertainty reductions for urban centers should not be too surprising: since the prior uncertainty is proportional the prior map of the emissions, and since in EDGAR urban emissions are high, the inversion will naturally apply large corrections and derive large uncertainty reductions for urban centers when using EDGAR as a prior. Does it yield urban emissions that are similar to that when using MAIOLICA as a prior ?

Yes. The posterior emission distribution in urban areas is very similar for both S-E and base inversion.

- l. 712: rewrite "model/observation pairs of one site" ?

We reworded this sentence.

- l. 748-750: the text is quite confusing and could be improved. Discussing the influence areas of the different sites and linking it to specific patterns of the corrections could feed this discussion. It raises some doubts regarding the corrections in the North East of Switzerland in the base inversion (see the general comment about it).

We tried to clarify our argument here. As we outline in this section this kind of attribution error was exactly the reason to include the additional sites JFJ and SSL in the study, because they show a largely different sensitivity pattern as compared to LAE and BEO.

- l. 755: it will be critical to strongly support the assumption that the corrections driven by GIM are erroneous since biased by a high signature of the local emissions at this site. Otherwise one could also think that the "shadowing effect" impacted the results when removing GIM and that the best estimate of the national emissions from the inversions should be obtained when using all the data.

L 755 does not discuss GIM but the LAE only case.

L760 discusses the GIM inversion. As discussed here and earlier in the text, we strongly believe that a combination of large local emissions and potential biases of the model in vertical mixing act together and seem to introduce a general low bias in our model simulations. Our method of estimating the model/observation uncertainty partly accounts for this misfit by estimating larger uncertainties for the site GIM. However, a general model bias cannot be treated by a Gaussian uncertainty alone and therefore we decided to exclude the site from the base inversion.

- l. 765: it is a dangerous discussion; it sounds like paradoxical to analyze the theoretical positive uncertainty reduction brought by GIM and FRU and to use it to demonstrate that they only have a local footprint, while it was stated earlier that assimilating these data increase significantly the errors on national budgets

The uncertainty discussed here represents only the analytical uncertainty, whereas the statement that these observations bias the emission estimate was based on more general considerations and the failure of the inversion system for these locally influenced sites. What we wanted to express here is that these additional observations in S-O5 did not improve posterior uncertainties and that this is also partly due to their localised footprint. We agree that the statement was confusing and we reworded accordingly.

- l 770-774 and 995-999: this assumes that the best inversion case is the base inversion, which is contradicted somewhere else. Having similar results with one site only, despite the "shadowing effects" that have been mentioned earlier, is actually a bit preoccupying. You may have to be more careful about such a discussion. Figure 5 was already showing the dominant role of BEO. But on principle, other sites should have been necessary to ensure that the incoming ("baseline") CH4 from remote areas is well constrained and that gradients between BEO and these sites could be used to constrain the fluxes in the inversion domain without high attribution uncertainties.

Although we don't want to promote the base inversion as our best estimate, we varied the sensitivity inversions only for one aspect at a time. Given all sensitivity inversions with variable observations we still think the base inversion should be understood as the reference in the context of this paragraph. However, we reworded the sentence to being a little bit less optimistic of running an inversion with a single site.

- Table 5 and line 815: why don't you rather look at the inventory for the year 2013 ? your inversions apply to 2013; can you systematically use the same denomination for the SGHGI / FOEN / national inventory ?

As mentioned in the text (l811 to l816) by the time we started this analysis only the reporting for the year 2012 was available in the 2014 reporting. The 2015 reporting showed relatively large changes compared to previous reporting (as given in Table5) and also included an estimate for 2013, which varied little from the 2012 value reported in 2015.

The inventories referred to as SGHGI, FOEN, and national inventory are all the same. We were not very careful with the naming, but improved this in the revised manuscript, sticking to SGHGI.

- l. 787 : these expectations highly depend on the definition of the baseline. See the general discussion about it.

As explained above our definition of baseline is driven by the 4-day backward calculations used in the study. See discussion above.

- l 800-804: such a comparison is a bit unadapted; their problem sounds like too different from yours

Although their study focusses on another species, we still think that this comparison is valid as it puts our results in perspective and indicates that uncertainties associated with the choice of transport model may indeed be larger than what we could identify with our two FLEXPART versions.

- l 825-840: having optimistic theoretical posterior uncertainties is not surprising since they rely on assumptions that the errors have unbiased and gaussian distributions, and that the set-up of the error covariances in the inversion system is perfect. When using the ML method, the analytical computation of the prior and obs/model variances rely on such assumptions and on the optimization of a few parameters, which correspond to a simple approximation of the actual errors. l. 837-838 are unclear.

While this may not be surprising, we wanted to emphasize the fact again that analytical uncertainty obtained from an individual inversion will in most cases not be sufficient to characterise the posterior uncertainty.

We did not understand which part of l837-838 is unclear and did change them.

- l. 845-846: please clarify what you mean by "referred"

We replaced the word.

- l 865: does the inversion yield such a seasonal variations for all agricultural areas or are there large areas where this does not apply ?

The general seasonal cycle can be observed in all agricultural areas, even those in eastern Switzerland, where emission were increased. These increases were strongest during spring/summer as can be seen in Figure 6. We added this statement to the text.

- in the abstract and at line 969 you should remind that the value you provide from SGHGI is for 2012

We added this information in the text and abstract.

- l. 993: I missed the link with the biosphere CO2 fluxes; the problem of the inversion of CO2 bio fluxes is quite different from that of inverting CH4 emissions and such a link is rather weak.

The link is the transport model. If the transport model is insufficient for the complex terrain encountered in the study area, then any CO_2 flux inversion is bound to fail. However, we weakened our statement so that it does not suggest that we can easily apply the current technique to CO_2 as well.

Additional references

Gerbig, C., Lin, J. C., Munger, J. W., and Wofsy, S. C.: What can tracer observations in the continental boundary layer tell us about surface-atmosphere fluxes?, Atmos. Chem. Phys., 6, 539-554, doi: 10.5194/acp-6-539-2006, 2006.

Hiller, R. V., Bretscher, D., DelSontro, T., Diem, T., Eugster, W., Henneberger, R., Hobi, S., Hodson, E., Imer, D., Kreuzer, M., Künzle, T., Merbold, L., Niklaus, P. A., Rihm, B., Schellenberger, A., Schroth, M. H., Schubert, C. J., Siegrist, H., Stieger, J., Buchmann, N., and Brunner, D.: Anthropogenic and natural methane fluxes in Switzerland synthesized within a spatially explicit inventory, Biogeosciences, 11, 1941-1959, doi: 10.5194/bg-11-1941-2014, 2014.

Rödenbeck, C., Houweling, S., Gloor, M., and Heimann, M.: CO_2 flux history 1982–2001 inferred from atmospheric data using a global inversion of atmospheric transport, Atmos. Chem. Phys., 3, 1919-1964, doi: 10.5194/acp-3-1919-2003, 2003.

Thompson, R. L., and Stohl, A.: FLEXINVERT: an atmospheric Bayesian inversion framework for determining surface fluxes of trace species using an optimized grid, Geosci. Model Dev., 7, 2223-2242, doi: 10.5194/gmd-7-2223-2014, 2014.