Authors' response to anonymous referee #2

In the following, referee's comments are in *italic*, authors' responses in normal font, and references (page, line, figure, and table number) to the revised manuscript in **bold**. Please note that this paper was merged with the accompanying paper, following the referees' comments and with approval from the Topical Editor. The summary of this paper was included in the Supplementary Material of the accompanying paper.

Although such sensitivity tests are undoubtedly important, the authors need to be clearer about how the outcomes of this work are of benefit to the wider inverse modelling community. In order to make the work more generally applicable perhaps the authors could provide some comparison of the relative importance of the input parameters, through a global sensitivity analysis for example. As it stands, the paper attempts to provide a justification for a particular model set-up to be used in the companion paper, but I wonder whether this is enough to justify a paper of its own, or whether this information should rather be included as a supplement to the companion paper.

Overall, I found the manuscript to be a little vague on what the outcomes of the sensitivity tests are, with a focus on qualitative rather than quantitative discussion. The manuscript was let down slightly by a number of grammatical errors or poorly constructed sentences, which may be why the key messages of the work are not clearly conveyed.

We acknowledge that some of the experiments were specifically meant to test the configuration of this system, following a custom of designing and presenting a new inverse modelling framework. Following the excellent suggestions from the reviewer, we decided to merge this paper with the accompanying paper. We hope the reviewer agrees that the sensitivities to vertical transport, and the attempt to separately estimate biosphere and anthropogenic fluxes, presented in the revised manuscript, are sufficiently discussed and useful to others, given that the sparsity of systems that perform inverse modelling of CH_4 fluxes.

We apologize for the inconsistencies that arose as a consequence of the weak formulation that existed in the manuscript. In this revision, we tried to more carefully phrase our text, and have also had the full paper language edited by a native English speaker. Moreover, we tried to make our descriptions more clear using new labelling.

General comments:
The paper focuses on sensitivity tests of various model inputs and parameters, and selects two models as a consequence of these tests. However, all tests assume the same model-data mismatches (mdm), which raises two major issues:

1. Clearly the mdm values are another input to the inversion which will change the form of R and thus the cost-function minimization. The impact of the mdm values on posterior emissions has been examined many times before (e.g. Michalak et al., 2005; Trudinger et al., 2007), with the studies commenting on the importance of these error terms. It seems a little odd therefore that this crucial component of the inversion is ignored in the sensitivity tests, given the somewhat arbitrary nature of their assignment. Furthermore, the observation error correlation structure would also impact on the solution, but I was unable to find any discussion of this in the manuscript.

This is an excellent comment and suggestion. In this study, we chose mdm based on quality of the observations, site types (mbl, land, tower, etc), and transport model error from forward runs, and a previous study by Bruhwiler *et al.* (2014), which used a similar system. Although the choice of the mdm values was somewhat arbitrary, they were generally targeted to have posterior Chi-squred values (Michalak *et al.*, 2005) spread around 1 (see also response to second referee of the accompanying paper on Point 3 of Scientific concerns). From the Chi-squared statistics, we found that the chosen mdm values were within the expectations to some extent. We will continuously develop the method to choose mdm values, but we hope the reviewer accepts the current choices.

For observation error correlation, we assumed all observations are independent of each other. However, it is known that observations are spatially and temporally correlated to some extent. The spatial correlation could be accounted for by considering e.g. distance between the sites. To take temporal correlation into account, further development is needed on the propagation of the observation covariance matrix. In the revised manuscript, we elaborated these issues further.

See e.g. Pg. 11 line 9-11 of the accompanying paper.

2. The decision to select the 2 chosen models S1 and S5 appears to be dependent on the posterior mismatch to the *observations. However, given this posterior is itself dependent on the chosen mdm, it is conceivable that under a different set of assumptions one would select a different model instead of S1 or S5. Since the values of mdm appear to be entirely down to investigator choice, I cannot see how the paper can propose an "optimal" inversion system that would be applicable beyond the specific case examined here.*

In the introduction it is stated that the aim of the paper is to "introduce the set-up. . .for an optimally working methane inversion system." However, I am not convinced that this is achievable from only seven different inversion configurations. Comparatively, one configuration may be better than the other six, but it would require a much more in-depth analysis to find the "optimum" configuration. For instance, some combination of configurations S2, S4 and S7 could provide a better match to the observations. However, performing the sensitivity analysis in the localised way of this work means that such a conclusion cannot be reached.

It may be that the choice of S1 and S5 is justified but any clear evidence to support this conclusion was either lost in the text or not present. In fact, Figure 7 would appear to suggest that there is very little to distinguish between the majority of configurations at both the global and continental scale.

> We agree with the reviewer that the choice of the set-up was based on expert knowledge rather than quantitative analysis. Also, the chosen set-up would not be optimal for the system. To find an optimal system, we need a much more extensive sensitivity analysis, including that of mdm, although we may not find an optimal set-up even after that. In this study, we did not find specific reasons to choose other set-ups than S1 and S5 based on uncertainty estimates and agreement with the NOAA in situ observations. Therefore, in the new manuscript, we decided to present each estimate as an equal realization of the surface fluxes. We hope the reviewer agrees that this is a more balanced representation of our results.

Specific comments:

Page 7, Lines 5-7: What was this "information provided by the experimentalists"? How many continuous day or night time observations were assimilated per site per day? What was the form of the observation error covariance matrix (e.g. diagonal?) If it is diagonal, is this assumption justified? In general some key details appear to be missing.

> For the selection of background observations, observation flags (measurement quality and assessment of background) provided by contributors were used. For example, for NOAA observations, observations without obvious problems during collection or analysis were chosen. Daily means from the selected observations were used in the system, and therefore, the number of observations per day was one.
>
> For the observations error covariance matrix, we did not assume any correlation between the observations, i.e. the matrix was diagonal. However, it is likely that some observations are temporally and spatially correlated. (see also above response to General comments 1).
>
> **Additional information was added in the accompanying paper. See Pg 8, line 14-15, 27-29.**

Page 7, Lines 14-16: Given the TM5 model is run at higher resolution over Europe, one might assume this would lead to a reduced representation error for European stations, i.e. 1x1 degree boxes might be able to represent a point in space better than a 4x6 degree. However, the mdm values appear to be the same whatever the model resolution at each site. Is there a reason for this?

> This is a very interesting point. We agree with the reviewer that resolving at higher transport model resolution generally reduces transport model error as it resolves the meteorological parameters and atmospheric mixing better, and also reduces "spreading" of emissions over large boxes (smearing). However, sites in Europe are also notoriously difficult to model due to their locations and uncertainty in emission sources. Furthermore, mdm also includes natural variability in measurements. Sites with small transport model error may have large mdm, if measurement variability is high. Therefore, we find that the mdm cannot be defined only based on the transport model resolution, and left the mdm of European sites not exceptionally small.

Page 7, Line 17: ". . .for the sites that appeared problematic in the inversions. . ." What meant by "appeared problematic"? Tuning the mdm values post-inversion is surely unacceptable as a violation of Bayes rule. What is the justification for 75 ppb? If the data is problematic why not just discard it completely?

> We agree with the reviewer that the mdm values should be chosen before inversion. In this study, the mdm was defined before inversion mainly based on quality of the observations, site types (mbl, land, tower, etc), and

transport model error from forward runs, and from a previous study by Bruhwiler *et al.* (2014), which used a similar system. From those, we learned which sites the model can, or can not, properly represent.

We acknowledge the concern of the reviewer about the effects of those sites with high mdm. It is indeed questionable how much information those observations provide to constrain the emissions. However, we did not simply remove them because we believed that some information could be useful. For example, for regions where observation network is sparse, the observation network is sparse, our emission estimates would act as a compensating effect by removing those observations, which may or may not be supported by the observations. Although we did not test thoroughly the effect of those observations by e.g. removing them, we hope the reviewer approves the use of these observations in the study.

Page 14, Lines 1-3: "S3 posterior mole fractions matched the observations best. . .in other words the additional information from the continuous observations was useful in gaining better agreement with NOAA observations." But according to Table 2, S3 is the configuration with discrete observations only, so the above statement cannot be right. It doesn't seem very surprising that S3 matches the NOAA observations better when these are the observations that have been used to constrain the emissions. Surely it would be more helpful to compare to an independent dataset rather than those that have already been used to derive the emissions.

Page 14, Lines 24-25: "Indeed the additional observation uncertainty increased the emissions uncertainty also." I fail to see how increasing the number of data points (however uncertain) would increase uncertainty, and this is not backed up by Table 4 which shows the inversion with only discrete observations (S3) has the highest emissions uncertainty.

Response to the above two comments on page 14:

We apologize for the misunderstanding and confusion that arose by poor phrasing. The paragraph was revised and included in the Supplementary Material of the merged paper as follows:

Removal of continuous observations decreased mean posterior anthropogenic emissions by about 70% in temperate North America and in southwest and east Europe. The decrease was partially compensated by an increase in biospheric emissions; for the North American temperate region, posterior biospheric emissions were about 100% larger without assimilating continuous observations, and the estimates were similar to the prior. Furthermore, the decrease was also compensated by >50% increase Asian tropic emission estimates. However, differences in biospheric emissions in the Asian temperate region were small. The reason could be that the discrete observations may have had little effect on the biospheric emissions, as the observations were located near anthropogenic sources. Therefore, the inversion less sensitive to biospheric emissions when continuous measurements are not assimilated. The effect of removing continuous observations was also significant in the uncertainty estimates, which were larger for anthropogenic emissions than for biospheric emissions. The posterior uncertainty for global anthropogenic emissions was about two times larger in the inversion not assimilating continuous observations, and the largest differences were found in the North American temperate and Asian temperate regions, and in southwest Europe. The posterior biospheric emission uncertainty was about three times larger in North American boreal, about twice as large in Asian temperate, and about 20% larger in North American temperate, Eurasian boreal, and Asian tropical regions than the estimates using continuous observations. These results indicate that improving prior estimates is important, especially for regions where observations are sparse.

Page 16, Line 7-8: "Thus, improving the prior estimates is important even when using an inverse model in the absence of observations." I assume this is a case of poor phrasing, but I am intrigued as to how one would perform inverse modelling in the absence of observations.

The sentence meant that, improving prior estimates is important especially for regions where the observation network is sparse. The phrase is revised and included in the Supplementary Material of the new manuscript.

Text revised: Pg . 5, line 17-18 of Supplementary Material of the accompanying paper.

Figure 2: I appreciate the Asian tropical plot is supposed to show the large variability with an ensemble size of 20, but is there a way of conveying the same message without it looking quite so messy? It might make the plot a little easier to interpret.

Here, our focus was to illustrate the variability of weekly estimates between random draws. Simply showing e.g. a range of estimates could mislead how estimates vary between weeks. To better illustrate the differences between E20 and E500, the estimates of E500 were plotted in thicker lines, although the we decided to retain the look.

See Fig S1 in Supplementary material of the accompanying paper.