Reply to referee#1

The authors would like to thank the referee for the careful review and the helpful comments. In the following, the reviewer's comments will be in **bold** font, and the responses will be in plain font, with suggested new text in *italics*.

On the gridded emissions. Some issues could be addressed to make things more robust and clear. It is indeed striking the visual differences in different approaches to gridding the EDGAR emissions in Figure 5. It would have helped me if you had mentioned in this section the spatial distribution of the native EDGAR inventory estimates and, how consistent country totals are after this gridding by the different methods (shown in Figure 11). Given the rather significant and arbitrary variations in the priors, a discussion of emission updates (Figures 6-8) becomes one that is related to two factors: the arbitrary errors in the priors because of the imperfect gridding process, and differences in model performance. At this point in the text only the second influence is considered, though it seems necessary to consider how the first factor is influencing the results too. (In other words, if all the models performed exactly the same in their inversion, there would still be substantially different updates apparent in Figure 6-8 because of the different gridding errors associated with the prior.) The better discussion of these issues comes later in the text in the comparison of figures 11 and 12, in my opinion. The authors might consider shortening or revising this earlier section.

Although the different grids have largely different resolutions and structure, all gridding algorithms are conserving the mass emitted in the original EDGAR v4.2 inventory. In that sense there are no "imperfections" or "gridding errors". Differences in the a priori emissions only occur for smaller spatial aggregates such as country totals, that do not perfectly align with the grid structure. We added this information as well as the spatial resolution of EDGAR v4.2. as follows:

*Although based on exactly the same EDGAR v4.2 inventory data, w**hich has a resolution of 0.1° x 0.1°**, the spatial aggregation to the different inversion grids leads to visually quite different distributions **despite the fact that all gridding algorithms are mass conserving, i.e. the emission from a coarse grid cell exactly corresponds to the sum of emissions from all finer EDGAR grid cells within that cell**.*

In this section, we tried to focus on the broad spatial patterns, which should be much less sensitive to the specific grid configuration than the analysis of country totals.

Regarding the apparent large differences in adjustments by the different models despite the reasonable similarity in posterior mole fraction time series generated by these models: It would seem that these aren't directly relatable unless you consider the sum of the fluxes shown in Figures 5 and 6, given that the posterior mixing ratios are from the sum of the prior emissions plus adjustments. Given the large apparent differences in the priors because of the different gridding approaches, this seems important to consider.

The mole fraction simulated for a given measurement location and time is determined by the fluxes within its footprint plus background. Assuming that the footprints of the transport models are similar/identical (which is certainly true for the three FLEXPART systems), a similarity in the time series can be translated into the expectation that the spatial emission patterns are similar, too. We agree that the fluxes correspond to the sum of Figures 5 and 6, but since there are no biases in the priors due to the conservation of emissions in each grid cell (as explained above), we think that Figure 6 alone is sufficient to discuss the broad spatial patterns.

On background levels. Since the approach for deriving background mole fractions taken by NILU is unique because it involves a subtraction related to the calculated influence of regional emissions on the observations deemed to represent background, it would seem reasonable to suggest that this subtraction might be causing the lower background mole fractions it derives. Is it not fairly easy to determine if this is the source of the offset?

The procedure of NILU indeed leads to a lower background as compared to the other approaches. Combined with the fact that NILU does not adjust this background in the inversion, this likely leads to comparatively high emissions. We will conduct another simulation with the EMPA2 system mimicking this approach. We don't expect, however, that this will explain all differences, because the difference between background with and without correction for regional influence is expected to be small. Nevertheless, this is a very valid suggestion that will be included as an additional sensitivity test in the revised manuscript.

Another minor issue, with regard to backgrounds for the approaches by EMPA2. The REBS approach is mentioned and an optimization is also indicated. Details about the optimization are lacking. Was the optimization applied to the REBS results? And how was that process constrained? Does the text mentioning that "the background is then allowed to evolve slowly with time" mean that it was just another optimized parameter in the inversion who's only constraint was low-frequency variation?

Indeed, EMPA2 optimized the REBS background levels. We will make this clear in the text as follows: *EMPA2 optimized the **REBS** background levels separately for each measurement site at selected reference points every 14 days. Background levels in between these reference points were linearly interpolated.*

And yes, in the EMPA system, which is sequentially applied to the data, the background level is another optimized parameter. It's update equation from one time step to the next follows the same equation (eq. 7) as the update for the emissions. The magnitude of the update uncertainty (ε_k) determines, how "slowly" the background is allowed to change from one time step to the next. We will add a reference to Equation (7) near the end of Sect. 2.3.

On section 3.3., uncertainty reductions. The authors seem to succeed in showing evidence refuting the initial statement that this is "a useful diagnostic" since the magnitudes seem primarily dependent on what is assumed as the uncertainty on the prior! In looking for robust conclusions from this section, there is one that I struggle to reconcile: How can uncertainty reductions expressed relative to absolute emission magnitudes be larger for those regions with higher emissions? Some explanation would be helpful here, since it seems not an expected or straightforward conclusion.

We fully agree that the discussion of uncertainty reductions is challenged by the fact that these strongly depend on the prior uncertainties. This issue is already addressed by the statement "Together with the different spatial uncertainty correlations, these differences have a marked effect on the resulting uncertainty reductions". We will better emphasize this issue already at the start of the section with a cautionary note:

However, it should be noted that the uncertainty reduction depends on the magnitude and correlation structure of the prior uncertainties. Comparing the uncertainty reductions thus helps illustrating the effect of the different model choices.

Details: Sentence two of abstract, consider adding a word: "but *emissions* have large GWPs and are, therefore..." Also, in the abstract the discrepancy in HFC-125 emissions estimated for the Iberian peninsula is the first point made in the comparison of results vs the UNFCCC inventory emissions, yet the main text mentions that "emissions from the Iberian countries are not well constrained by the current observation network." Some modifications to the abstract seem necessary.

We don't think that adding "emissions" would make the sentence more easily understandable. It is common practice to refer to the GWP of a gas rather than to the GWP of its emissions. Ultimately, it is the gas itself that has the properties leading to a high or low GWP.

It is true that emissions are not very well constrained for the Iberian Peninsula. Nevertheless, the fact that all models estimate much higher than reported emissions for HFC-125 but not for HFC-134a, is a strong indication that HFC-125 emissions are underreported. We will add a note of caution to the abstract:

.. though with a large scatter between individual estimates.

Define "standard deviation (normalized)" in the caption of the figure showing Taylor diagrams. I presume it is the ratio between the observed vs posterior calculated mole fractions. Tthis should be mentioned if true. Any de-trending applied to the results over the year, or is it just the s.d. of the annual data record considered together?

The word "normalized" refers to the fact that in a Taylor diagram the standard deviation of the simulated values is normalized by the standard deviation of the observations. A value of 1 indicates perfect agreement between the magnitude of scatter in the simulated and observed values. This information will be added to the caption.

Figure 1 caption, mention that the reduced grid is only associated with the EMPA simulations, if true.

Correct, the figure caption was indeed lacking and will be change to:

Annual mean surface sensitivity in units of [ppb/(kg m-2 s-1)] for (a) the original 0.1°x0.1°grid and (b) for the reduced grid of the FLEXPART-based model system EMPA.