Reply to referee#2

The authors would like to thank anonymous referee #2 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*.

The conclusions on country-wide emissions appear somewhat unconsolidated given that the model-to-model differences are as large as the estimated emissions for some countries (e.g. Figure 12). While I accept the approach to use model versions that are as close as possible to the respective production settings, it is quite unsatisfying that the reasons for these model differences are essentially unresolved. In that context, I am also not convinced by using the model median value (of only 4 models). I would suggest making abstract and conclusions somewhat humble by adding some more discussion on how the discrepancies between bottom-up and top-down emissions compare to model differences.

Being humble in terms of conclusions about country scale emissions is a valid suggestion. In the abstract we will expand the sentence regarding the much higher simulated than reported HFC-125 emissions from Spain+Portugal with

.. though with a large scatter between individual estimates

and will add "country-scale" to the last sentence to read as follows:

*.. but a denser network would be needed for more reliable monitoring of **country-scale** emissions of these important greenhouse gases across Europe.*

In the conclusions section, the limitations of the inversions with respect to country emissions were already pointed out quite clearly, e.g. in the third last paragraph with the sentences

"*However, the estimates of the individual models varied considerably. Considering all three gases and the largest countries, the scatter was smallest for the UK (1σ standard deviation of 3-11%), followed by France (8-15%), Germany (19-22%), Italy (12-31%), and Spain+Portugal (24-30%). The individual models often did not overlap within the range of the combined uncertainties suggesting that ..*"

and in the last paragraph with

"*The network has the potential to identify significant shortcomings in the nationally reported emissions but a denser network would be needed for a more accurate assignment to individual countries. Model-to-model differences were often very large whereas the model median appears to have significant skill as judged from the comparison with reported HFC-134a emissions, which are considered to be relatively well known*."

Nevertheless, to better emphasize the wide range of country estimates, we will replace the standard uncertainties of the means by the standard deviations of the individual estimates (in percent of the mean) and add another sentence on typical ranges between minimum and maximum. The sentences in the 3rd last paragraph will read as follows:

Considering all three gases and the largest countries and defining "scatter" by the 1σ standard deviation of individual estimates (in % of the mean), the scatter was smallest for the UK (5-22%), followed by France (16-28%), Germany (38-43%), Italy (23-63%), and Spain+Portugal (42-51%). Differences between minimum and maximum estimates for a given country were often as large as a factor 2, sometimes even a factor of 3, especially for Italy and Spain+Portugal.

Furthermore, the last sentence in the conclusions will be changed to

*Model-to-model differences were often very large, **occasionally as large as the estimated emissions**, whereas the median appears to ..*

It is difficult to provide a useful statistics summarizing the results of an ensemble of only 4 models. Nevertheless, the median is more robust than the mean value and is commonly used for model ensembles. Note that we also show the full range of the model estimates (in Fig. 13), not only the medians.

A detail that came to my attention is that the release height for the particles at Jungfraujoch was adjusted for the NAME model to match the FLEXPART footprints. Essentially, this adjustment appears arbitrary and contradicts the general philosophy to use production settings for each model. If the adjustment was not made (transport induced) model differences would be even larger. So, given that (at least one of) the transport models are not able to correctly model transport at the mountain sites, how confident are you with respect to your overall conclusions?

Unlike for FLEXPART, we did not do any independent analysis on the best release height for the NAME model. Previous analysis provided an optimum release height for FLEXPART. Instead, we used a release height for NAME that produced model time series as close as possible to FLEXPART's given a specific emissions field. We will explain this in the text. This approach allowed us to include the results of NAME despite of the difficulties in representing this mountain site. A thorough investigation of the reasons for the differences between FLEXPART and NAME for Jungfraujoch would be desirable, but was not feasible within the scope of this project.

P2,L24: regulated reported -> reported

Done

P9,L6 and following: Occasionally, I got confused by the naming conventions. I would suggest using NAME and FLEXPART when referring to transport issues and the others names when referring to the entire modelling systems: P9,L6: UKMO -> NAME, P9,L13: NAME->UKMO, check other places.

We changed the sentence that confused the reviewer to:

In particular, the score of the NAME-based system UKMO is moving closer to the three FLEXPART-based systems EMPA, EMPA2, and NILU.