The authors used four inverse models to estimate European emissions of HFC-134a, HFC-125 and SF_6 for the year 2011. All systems used measurements from Jungfraujoch, Mace Head, and Monte Cimone. The paper is well written and provides interesting insights. I think the main problem of the paper was that the differences in the choices, such as spatial correlations of the prior and background treatment, had a quite substantial impact on differences among the models. What was the reason that those were not controlled? If they were better controlled, maybe we could have had more insights on which models are doing better and what we might do to improve the emissions estimation through inverse modeling. Below are some other comments and questions I had and I would recommend publication after they are addressed.

For Figure 1, is this the sensitivity created using FLEXPART or NAME? I would also assume that the sensitivity is quite different depending on the month. Which month is this? And is this the monthly mean?

It was a little unclear why NAME needed such a high release height at Jungfraujoch. If the point of the paper is to better understand the differences among the four inversion systems, I find it puzzling that the authors would modify to make the model footprint sensitivities comparable to each other.

I had a hard time understanding how the emissions were created following the country outlines. What was the means used to split the EDGAR grid to country outlines? Also, because the prior emissions are so different, I find it more informative if the Fig. 6 was not comparing between prior and posterior but EDGAR and posterior.

Why did EMPA2 use the uncertainty set uniformly to 137%? This seemed a little strange and was curious for the reason behind this specific value.

One of the explanations for why UK's estimated emissions are much higher than what is reported to UNFCCC, the authors mention the use of an assumed high loss rate of HFC-134a from car air conditioning systems in the UK. Why is this only in the UK and how different is the loss rate among the countries? Is a similar explanation possible for overestimation and/or underestimation for different species?

Backwards mode time differ substantially among the models and I would have expected UKMO to have larger difference between prior and posterior away from the measurement sites, compared to the other model systems that have shorter time span. Why is it that UKMO shows almost no difference between the two farther away from the measurement sites?

Minor comments

1. Sometimes authors state the country by name and sometimes by the ISO2 convention country code. It is a little confusing to me and so I would suggest to be consistent and I would appreciate if there was a table listing the country names with ISO2 code if the authors want to use the codes.

2. P. 11 l. 4 “An important question is the context … is the question” → delete the second “the question” in the sentence to make it “… Paris Agreement is, how suitable is…”
3. I am not quite sure what 0.1°x0.1°min means in Table 1.
4. “reduced acc. to” → “reduced according to” in Table 1 for UKMO
5. State vector length is mentioned in Table 1 but was not explained in the text at all. Can this be clarified in terms of how this is used in the equation and why the equations look so different depending on the system?
6. How is the EDGAR prior uncertainty determined in Figure 13? I find that to be a little misleading, since I do not think EDGAR provides such a value.
7. Figure 14 is very difficult to see – maybe a different color scheme would work better.