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, ***which 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.

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.

Reply to referee#3

The authors would like to thank anonymous referee #3 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 authors used four inverse models to estimate European emissions of HFC-134a, HFC-125 and SF6 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.

The main motivation was to document the uncertainty associated with the different choices that have been made in recent halocarbon inversion studies. There is no doubt that differences would have been substantially smaller with a more strongly controlled setup. It was not our intention to assess the quality of the transport simulations of FLEXPART as compared to NAME, which it would ultimately come down to if all other choices were identical.

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?

We apologize that the figure caption was not sufficiently clear (as also noted by another reviewer). It will be changed 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.

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.

As also mentioned in our reply to reviewer #2, the measurements from Jungfraujoch have not been used in previous inversion studies based on the NAME model, because the results had not been satisfactory for this site and no independent analysis on the optimal release height had been conducted before, in contrast to FLEXPART. The approach chosen here was pragmatic, so as to not disadvantage the NAME model, and allowed us to include the results of NAME despite of these difficulties. 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.

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.**

We will add the information that the original resolution of EDGAR was 0.1° x 0.1°. In the case of the UKMO system, EDGAR emissions were first regridded to a fine grid, and the country outlines were then followed as closely as possible. Each grid cell was assigned to the country with the largest share. Except for small countries, the error introduced by this procedure with cells at the borders shared by more than one country is small. We extended the sentence explaining the grid for the UKMO system with

follows the country outlines ***as closely as possible given the resolution of a fine grid uncerlying the reduced inversion grid***.

EDGAR is the prior. Note that the prior emissions are not different, only their spatial representation. Comparing the results with EDGAR at the original resolution would require redistributing the emissions estimated on the reduced grid to the original fine grid. This would be doable technically, but it would give a wrong impression of high resolution of the inverted emissions. We strongly prefer the present representation of the results.

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.

The 137% was a result of the requirement, that the total uncertainty of a domain covering most of Europe was 20%. We will add this information to the text.

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?

The UK inventory is conservative (overestimates), as it assumes that there is a 100% replacement of air conditioning fluid in all mobile air conditioning systems each year. Each country makes their own choice in this aspect provided it is backed by expert knowledge. It is not clear what every country across Europe does in this respect. This particular situation is specific for HFC-134a but other issues will undoubtedly impact the emissions of different countries for different gases.

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?

A backward simulation over 5 days captures most of the sensitivities of the measurements to emissions within Europe because the sensitivity decreases very rapidly as time and distance from the measurement increases. Extending to 10 days (NILU) or even 19 days (UKMO) changes little. The fact that UKMO adjusts relatively little at larger distances from the sites must be due to the specific choices of a priori versus observation uncertainties.

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.

Instead of adding another table we added the country names in the caption of Figure 11:

CH=Switzerland, DE=Germany, IT=Italy, FR=France, ES=Spain, PT=Portugal, UK=United Kingdom, IR=Ireland, BE=Belgium, NL=Netherlands, LU=Luxemburg, AT=Austria, DK=Denmark, SW=Sweden, FI=Finland, PO=Poland, CZ=Czech Republic, SV=Slovakia, NO=Norway.

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…"

Thank you, done

3. I am not quite sure what 0.1°x0.1°min means in Table 1.

It means that the minimum size of a grid cell is 0.1°x0.1°, but larger when cells are aggregated for the reduced grid. We will change "min." to "minimum".

4. “reduced acc. to” -> “reduced according to” in Table 1 for UKMO

Done

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?

We don't quite agree: The state vector ***x*** was introduced in the context of equation 1 as follows: "***x*** *is the state vector which includes the gridded emissions and possibly other elements such as background mole fractions, and n is the number of state vector elements to be estimated/optimized by the inversion*." The state vector was also referred to at other locations, e.g. in the sentence "*In the EMPA system, a single element per observation site is added to the state vector to represent the background at time step k.*"

To better link the information in Table 1 with the text we will add the following line after the first sentence mentioned above:

An overview of the number and type of state vector elements used in each system is provided in Table 1.

Note that the equations presented in the manuscript do not depend on the specific choices of state vector elements.

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.

EDGAR indeed does not report uncertainties. These uncertainties denote the range of uncertainties in the prior that is introduced by the different gridding methods. This information will be added to the caption.

7. Figure 14 is very difficult to see – maybe a different color scheme would work better.

We played a lot with different color schemes and found that using a single color in different shadings works best. The main issue is that the different grid shapes already introduce a lot of variation, such that the figures become too complex and less easily readable when using a color scheme composed of multiple colors.

Comparison of four inverse modelling systems applied to the estimation of HFC-125, HFC-134a and SF_6 emissions over Europe

Dominik Brunner[1], Tim Arnold[2,3,4], Stephan Henne[1], Alistair Manning[2], Rona L. Thompson[5], Michela Maione[6], Simon O'Doherty[7], and Stefan Reimann[1]

[1]Laboratory for Air Pollution/Environmental Technology, Empa, Swiss Federal Laboratories for Materials Science and Technology, Dübendorf, 8600, Switzerland
[2]Met Office, Exeter EX1 3PB, United Kingdom
[3]National Physical Laboratory, Teddington, Middlesex TW11 0LW, UK
[4]School of GeoSciences, University of Edinburgh, Edinburgh, EH9 3FF, UK
[5]NILU - Norwegian Institute for Air Research, Kjeller, 2007, Norway
[6]Dipartimento di Scienze Pure e Applicate (DiSPeA), University of Urbino "Carlo Bo", Urbino, 61029, Italy
[7]School of Chemistry, University of Bristol, Bristol, UK

Correspondence to: Dominik Brunner (dominik.brunner@empa.ch)

Abstract. Hydrofluorocarbons (HFCs) are used in a range of industrial applications and have largely replaced previously used gases (CFCs and HCFCs). HFCs are not ozone depleting but have large global warming potentials and are, therefore, reported to the United Nations Framework Convention on Climate Change (UNFCCC). Here, we use four independent inverse models to estimate European emissions of the two HFCs contributing the most to global warming (HFC-134a and HFC-125) and of SF_6 for the year 2011. Using an ensemble of inverse models offers the possibility to better understand systematic uncertainties in inversions. All systems relied on the same measurement time series from Jungfraujoch (Switzerland), Mace Head (Ireland), and Monte Cimone (Italy), and the same a priori emissions, but differed in terms of Lagrangian transport model (FLEXPART, NAME), inversion method (Bayesian, Extended Kalman Filter), treatment of background mole fractions, spatial gridding, and a priori uncertainties. The model systems were compared with respect to the ability to reproduce the measurement time series, the spatial distribution of the posterior emissions, uncertainty reductions, and total emissions estimated for selected countries. All systems were able to reproduce the measurement time series very well with prior correlations between 0.5 and 0.9 and posterior correlations being higher by 0.05 to 0.1. For HFC-125, all models estimated higher emissions from Spain+Portugal than reported to UNFCCC (median higher by 390%) though with a large scatter between individual estimates. Estimates for Germany (+140%) and Ireland (+850%) were also considerably higher than UNFCCC, whereas the estimates for France and the UK were consistent with the national reports. In contrast to HFC-125, HFC-134a emissions from Spain+Portugal were broadly consistent with UNFCCC, and emissions from Germany were only 30% higher. The data suggests that the UK over reports its HFC-134a emissions to UNFCCC, as the model median emission was significantly lower, by 50%. An overestimation of both HFC-125 and HFC-134a emissions by about a factor 2 was also found for a group of eastern European countries (Czech Republic + Poland + Slovakia), though with less confidence since the measurement network has a low sensitivity to these countries. Consistent with UNFCCC, the models identified Germany as the highest national emitter of SF_6 in Europe, and the model median emission was only 1% lower than the UNFCCC numbers. In contrast, the model median emissions were 2-3 times higher than UNFCCC numbers for Italy, France and Spain+Portugal. The country-aggregated emissions from the different models often did not overlap within the range of the analytical uncertainties formally given by the inversion systems, suggesting that parametric and structural uncertainties are often dominant in the overall a posteriori uncertainty. The current European network of three routine monitoring sites for synthetic greenhouse gases has the potential to identify significant shortcomings in nationally reported

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

1 Introduction

Synthetic halocarbons are used for a wide range of applications such as refrigeration and air conditioning, foams, solvents, aerosol products and fire protection. The first generation of compounds, the chlorine containing chlorofluorocarbons (CFCs) and bromine containing halons, were harmful to the stratospheric ozone layer and were phased-out under the Montreal Protocol entering into force in 1987. They were substituted by natural refrigerants including hydrocarbons and ammonia and by another class of halocarbons, the hydro-chlorofluorocarbons (HCFCs), which have lower stratospheric ozone depletion potentials (ODPs) and lower global warming potentials (GWPs) than the CFCs. Regulation of the production and consumption of HCFCs under the Montreal Protocol led to a strong decline in their emissions over Europe after 2004 (Brunner et al., 2012; Derwent et al., 2007; Graziosi et al., 2015) whereas emissions were still increasing in developing countries until recently (Saikawa et al., 2012; Xiang et al., 2014). Today, HCFCs and CFCs are mainly replaced by chlorine-free hydrofluorocarbons (HFCs), which are no longer harmful to the ozone layer except for minor indirect effects (Hurwitz et al., 2015), but some have large GWPs.

Current emissions of HFCs and CFCs are equivalent to only about 5% of global CO_2 emissions on a CO_2-equivalent basis, but, as Velders et al. (2009) highlighted, in a business as usual scenario without further regulations, HFC emissions could grow to an equivalent of 9 – 19% of projected global CO_2 emissions by 2050, stressing the need for binding emission regulations. In view of the urgency of the problem and the success of the Paris Agreement, 197 countries adopted in October 2016 an amendment to the Montreal Protocol to phase down the emissions of HFCs by more than 80% over the next 30 years.

HFC-134a and HFC-125 considered in this study are the two most abundant HFCs in Europe constituting 69% of all HFC emissions (CO_2-eq.) in 2012, with HFC-143a contributing another 23% according to officially reported emissions of the EU-28 countries. HFC-134a has a 100-yr GWP of 1,300 and is the preferred refrigerant in motor vehicle air conditioning systems. HFC-125 has a GWP of 3,170 and is mainly used in refrigerant blends for residential and commercial refrigeration and in smaller amounts as a fire suppression agent (O'Doherty et al., 2009; Velders et al., 2009). Sulfur hexafluoride (SF_6) is primarily used as a dielectric and insulator in high-voltage electronic installations. With a GWP of around 22,800, SF_6 is the most potent greenhouse gas reported to UNFCCC. SF_6 emissions are equivalent to about 0.5% of current global CO_2 emissions (CO_2-eq.), but emissions are still growing, especially in developing countries (Levin et al., 2010; Rigby et al., 2010).

Due to their long atmospheric lifetime, HFCs and SF_6 are rather uniformly distributed in the atmosphere. Global emissions can, therefore, be estimated from measurements at a few representative baseline stations distributed across the globe (Cunnold et al., 1994; Montzka et al., 2015; Vollmer et al., 2011; Xiang et al., 2014). Estimating emissions at continental or even regional and country scale, however, requires a denser network of sites with varying sensitivity to emissions from the region of interest (Villani et al., 2010).

Currently, HFCs are routinely measured at only three sites in Europe: Jungfraujoch in Switzerland, Mace Head in Ireland, and Monte Cimone in Italy. Measurements from these sites have been used in several previous inverse modeling studies to estimate European emissions of selected halocarbons and SF_6 (Brunner et al., 2012; Ganesan et al., 2014; Keller et al., 2011; Keller et al., 2012; Lunt et al., 2015; Maione et al., 2014; Manning, 2011; Manning et al., 2003; Rigby et al., 2011; Simmonds et al., 2016; Stohl et al., 2009). Different Lagrangian transport models and inversion approaches have been applied in these studies but no systematic comparison between the model systems has been undertaken so far. The European infrastructure project InGOS (Integrated non-CO_2 Greenhouse gas Observation System) helped to improve the quality and compatibility of these measurements, to further develop the measurement technologies, and to collect and harmonize the data. It also supported a range of modeling studies to quantify European emissions of non-CO_2 greenhouse gases including CH_4 and N_2O (Bergamaschi et al., 2015) and halocarbons (this study), and to evaluate the models with respect to their transport properties.

Inverse emission estimation using direct atmospheric observations (commonly referred to as ‘top-down’) has been proposed as a tool for helping to verify anthropogenic emission inventories estimated by the individual countries based on statistical data and source-specific emission factors (commonly referred to as ‘bottom-up’) (Nisbet and Weiss, 2010). However, to enhance the credibility of this top-down approach, a better understanding of the associated uncertainties is needed. Currently, there is no commonly accepted benchmark against which to test the models and there is no single emission source that is known well enough to serve this purpose. Emissions of radon, for example, have turned out to be spatially and temporally more variable than previously thought (Karstens et al., 2015). Large-scale tracer release experiments such as ETEX (Van dop et al., 1998) have been instrumental in the development of dispersion models, but their temporal and spatial coverage is too sparse for an overall assessment of atmospheric transport and inverse modeling systems. Traditionally, inverse modeling studies have applied a single transport model and inversion setup and reported posterior uncertainties deduced from Gaussian error statistics in a Bayesian framework. More recently, awareness has grown that this approach may miss important contributions to the true uncertainties, including errors in model transport, representation errors, and uncertainties related to the chosen setup and the expert judgments that classical Bayesian inversions heavily rely on. Approaches to overcome these limitations included a better consideration of transport uncertainties (Baker et al., 2006; Lin and Gerbig, 2005; Locatelli et al., 2013), objective estimation of error covariance parameters (Berchet et al., 2013; Brunner et al., 2012; Michalak et al., 2005), and model experiments exploring the sensitivity of the results to different assumptions (Bergamaschi et al., 2010; Brunner et al., 2012; Henne et al., 2016) . A promising new avenue is to extend the classical Bayesian framework with the dimension of ‘uncertainties of uncertainties’ (Berchet et al., 2015; Ganesan et al., 2014).

Here we apply four independent inversion systems to quantify the emissions of HFC-134a, HFC-125 and SF_6 over Europe for the year 2011 in a set of well-defined model experiments with common observation data and a priori emissions. We aim to compare the results of four well-established systems used in previous studies and to better assess the uncertainties associated with different choices of transport model, inversion method, treatment of background mole fractions, spatial gridding, a priori uncertainties, and error correlation structures, which add to the analytical uncertainties determined by the individual systems. Furthermore, we aim to evaluate the ability of the current network of three monitoring sites in Europe to constrain the emissions of synthetic greenhouse gases in individual European countries.

2 Methods

2.1 Observation Data

Measurements were available as hourly or two-hourly samples from the coastal site, Mace Head (9.90°W, 53.33°N, 15 m above mean sea level (amsl)), Ireland, and the two mountain sites, Jungfraujoch (7.99°E, 46.55°N, 3573 m amsl), Switzerland, and Monte Cimone (10.70°E, 44.18°N, 2165 m amsl), Italy. Halocarbons and SF_6 are measured at Jungfraujoch and Mace Head with a 'Medusa' Gas Chromatography/Mass Spectrometry (GC/MS) system (Miller et al., 2008). At Monte Cimone, an Adsorption Desorption System (ADS) GC/MS (Maione et al., 2013) is used, which does not enable SF_6 to be measured. The measurement data and their uncertainties (1σ single measurement precision determined as running mean of calibration standards bracketing each measurement) were provided to all groups at their native time resolution. Typical precisions for HFC-134a, HFC-125 and SF_6 are in the range 0.2-0.5 ppt, 0.05-0.1 ppt and 0.02-0.03 ppt, respectively.

For the assimilation, these observations were averaged to 3-hourly values in the EMPA and EMPA2 models and to daily means in NILU. UKMO used a single 3-hourly mean value per day around the time when the uncertainty of boundary layer heights was considered to be lowest, i.e. in the early afternoon (12-15 UTC) at Mace Head, and when the least influence from local boundary layer transport can be expected at the two mountain sites (06-09 UTC).

2.2 Inverse Modelling Systems

A brief overview of the four inversion systems employed in this study is presented in Table 1. All systems have been used in similar configurations in previous studies as referenced in the table. In all systems, atmospheric transport was described by a Lagrangian Particle Dispersion Model (LPDM). The LPDMs were operated in backwards in time, receptor-oriented mode (Seibert and Frank, 2004). In this mode, virtual particles (infinitesimally small air parcels) are released at the measurement sites and followed backwards in time, typically for a few days.

Three systems (EMPA, EMPA2, NILU) used the transport model FLEXPART (Stohl et al., 2005) driven by 3-hourly analysis and forecast fields from the European Centre for Medium Range Weather Forecasts - Integrated Forecast System (ECMWF-IFS). The fourth system, UKMO, relied on the transport model NAME (Ryall and Maryon, 1998) driven by global analyses of the UK Met Office's Numerical Weather Prediction model.

The outputs of the LPDMs are emission sensitivity maps, so-called 'footprints', for each particle ensemble release time. The footprints represent the total sensitivity of an observation to surface emissions over the backwards simulation time. Multiplying the footprint by an emission map and integrating in space and time gives a simulated mole fraction at each release time and location. Assuming temporally constant emissions for the inversion period, the relation between emissions and simulated mole fractions can be written as

$$\mathbf{y} = \mathbf{Mx}, \tag{1}$$

where $\mathbf{y} = (y_1 \quad \cdots \quad y_m)$ is the vector of simulated mole fractions at all times and stations, with m being the total number of available measurements. $\mathbf{x} = (x_1 \quad \cdots \quad x_n)$ is the state vector which includes the gridded emissions and possibly other elements such as background mole fractions, and n is the number of state vector elements to be estimated/optimized by the inversion. An overview of the number and type of state vector elements used in each system is provided in Table 1. **M** is the sensitivity matrix (with dimension m x n),

$$\mathbf{M} = \begin{pmatrix} M_{1,1} & \cdots & M_{1,n} \\ \vdots & \ddots & \vdots \\ M_{m,1} & \cdots & M_{m,n} \end{pmatrix}. \tag{2}$$

Each row of **M** describes the sensitivity of a given measurement to all state vector elements composed of the footprint computed by the LPDM and possibly other elements such as the sensitivity to the background field (see e.g. Thompson and Stohl, 2014).

The goal of the inversion is to estimate an optimized state **x**, which accounts for the observed mole fractions $\mathbf{y}_o$ by reducing the difference between observed and simulated values, additionally constrained by the uncertainty bounds of the prior state variables. In the Bayesian framework and assuming Gaussian uncertainty distributions, this optimized state is obtained by minimizing the following cost function *J(***x***)* (e.g. Tarantola, 2005)

$$J(\mathbf{x}) = \frac{1}{2}(\mathbf{x} - \mathbf{x}_\mathrm{b})^\mathrm{T}\mathbf{B}^{-1}(\mathbf{x} - \mathbf{x}_\mathrm{b}) + \frac{1}{2}(\mathbf{Mx} - \mathbf{y}_\mathrm{o})^\mathrm{T}\mathbf{R}^{-1}(\mathbf{Mx} - \mathbf{y}_\mathrm{o}). \tag{3}$$

The first term on the right-hand side describes the deviation of the optimized state $\mathbf{x}$ from a prior state $\mathbf{x}_\mathrm{b}$, the second term the deviation of the simulated mole fractions from the observations. Both terms are weighted by their uncertainties represented by the error covariance matrices $\mathbf{B}$ (n x n) and $\mathbf{R}$ (m x m) for the prior and observation uncertainties, respectively.

This approach was employed by the inversion systems EMPA2, NILU and UKMO, which, however, differed in various other aspects of the implementation. In order to mimic the approach presented by Stohl et al. (2009) as closely as possible, EMPA2 assumed the matrices $\mathbf{B}$ and $\mathbf{R}$ to be diagonal (i.e., uncorrelated errors). NILU, instead, assumed a correlation length scale of 200 km over land and 1000 km over ocean for the prior emission field, and $\mathbf{R}$ contained off-diagonal elements to represent the cross-correlations of the model representation error (see Thompson and Stohl, 2014). Like EMPA2, UKMO did not account for potentially correlated errors in the prior emission field. As will be shown in Sect. 3, the choice of correlation structure has quite a strong influence on the results. Due to the way bottom-up inventories are generated, it may be justified to assume stronger error correlations within a country than across country borders, but none of the inversion systems adopted such a strategy.

To avoid non-physical negative emissions, NILU applied a 'truncated Gaussian' approach (Thacker, 2007; Thompson and Stohl, 2014). This entails performing a second step after the inversion in which an inequality constraint, namely that the emissions must be greater than or equal to zero, is applied accounting also for the error-covariance between grid-cells.

EMPA2 estimated the model uncertainty following the suggestions by Stohl et al. (2009). In a first step, the Root Mean Square Error (RMSE) of the prior simulation minus observations was calculated for each site separately. The model residuals were then scaled by the RMSE. The normalized residual distribution often does not follow a normal distribution, but is skewed towards large negative values (large model underestimations). In order to reduce the influence of such points in the inversion, the model uncertainty for these 'outliers' was iteratively adjusted so that the normalized residual distribution followed a normal distribution more closely. This procedure was repeated using the posterior simulations of a first inversion run. A second and third inversion run was then performed using the updated model uncertainties but the same prior state. Furthermore, prior uncertainties were reduced for grid cells with negative posterior emissions, and the inversion was iterated until a solution without significant negative emission contributions was obtained, again following the suggestion by Stohl et al. (2009).

The Met Office's inverse modelling system (InTEM – Inversion Technique for Emission Modelling) using the NAME model has evolved since the work of Manning et al. (2011) and the NitroEurope project (Bergamaschi et al., 2015) and is now based on a Bayesian methodology. Measurement uncertainty reported in the InGOS data set was used as observation error. Model-measurement mismatch errors were also applied to each measurement and were calculated using a metric based on the degree of influence of local fluxes on the measurement (Manning et al., 2011). These model errors were inflated based on the difference between the model release height above sea level and the true altitude of the observation, and the relative difference between the modelled boundary layer height and the observation height. No spatial or temporal correlations were applied in these inversions. Grid boxes were aggregated based on the sensitivity of measurements to emissions, creating around 100-150 course grid regions within the inversion domain. A non-negative least square solver was used to optimise the solution thus preventing negative emissions from being estimated.

EMPA applied an extended Kalman Filter (ExtKF) as described in detail in Brunner et al. (2012). Different from the other systems the observations are not used all at the same time, but are assimilated sequentially thereby gradually adjusting the state to a solution that is optimal given all past observations up to the assimilation time. The Kalman filter update equations are for the state:

$$\mathbf{x}_\mathrm{k}^{+} = \mathbf{x}_\mathrm{k}^{-} + \mathbf{K}_\mathrm{k}(\mathbf{y}_\mathrm{k} - \mathbf{M}_\mathrm{k}\mathbf{x}_\mathrm{k}^{-}) \quad (4)$$

and for the uncertainty of the state:

$$\mathbf{P}_\mathrm{k}^{+} = (\mathbf{1} - \mathbf{K}_\mathrm{k}\mathbf{M}_\mathrm{k})\mathbf{P}_\mathrm{k}^{-} \quad (5)$$

where k is the time index, $\mathbf{K}_\mathrm{k}$ the Kalman Gain matrix, defined as

$$\mathbf{K}_\mathrm{k} = \mathbf{P}_\mathrm{k}^{-}\mathbf{M}_\mathrm{k}^{\mathrm{T}}\left(\mathbf{R}_\mathrm{k} - \mathbf{M}_\mathrm{k}\mathbf{P}_\mathrm{k}^{-}\mathbf{M}_\mathrm{k}^{\mathrm{T}}\right)^{-1}, \quad (6)$$

$\mathbf{P}_\mathrm{k}$ the state error covariance matrix, and $\mathbf{M}_\mathrm{k}$ the sensitivity matrix for time k. The minus sign denotes a ‘first guess’ state before assimilation of the observation $\mathbf{y}_\mathrm{k}$ available at time k, and the plus sign denotes the ‘analysis’ state after assimilation. The matrix $\mathbf{P}$ essentially takes the role of $\mathbf{B}$ in the Bayesian inversion and the observation and model representation uncertainty matrix $\mathbf{R}$ is included in the definition of the Kalman Gain matrix. The similarity between the Kalman Filter and Bayesian inversion is further illustrated by the fact that the solution to Eq. (3) is given by the same Eq. (4) but with $\mathbf{B}$ replacing $\mathbf{P}_\mathbf{k}^{-}$ in the Kalman Gain matrix and all observations being used at once instead of looping over time steps k. Different from the Bayesian inversions, however, the emissions were not assumed to be constant but to evolve slowly with time as expressed by the forecast equation

$$\mathbf{x}_{\mathrm{k+1}}^{-} = \mathbf{x}_\mathrm{k}^{+} + \boldsymbol{\varepsilon}_\mathrm{k}, \quad (7)$$

which states that the emissions at time k+1 are expected to be the same as at time k within an uncertainty ε_k. This step adds uncertainty to the emissions according to

$$\mathbf{P}_{\mathrm{k+1}}^{-} = \mathbf{P}_\mathrm{k}^{+} + \mathbf{Q}_\mathrm{k}, \quad (8)$$

so that the uncertainty can grow with time in regions poorly covered by the observations. This is different from the other inversions where the posterior uncertainties are always smaller than the prior uncertainties. Without this forecast step, the solution after assimilating all observations would be identical to the solution obtained with Eq. (3). The new matrix $\mathbf{Q}_\mathrm{k}$, which has no correspondence in the Bayesian inversion, describes the uncertainty of the forecast and determines how rapidly the emissions (and background levels, see below) are allowed to change with time.

Another unique feature of the EMPA system is that it estimates the logarithm of the emissions in order to constrain the solution to positive values. This makes the problem non-linear and, therefore, requires the application of an Extended Kalman Filter that linearizes the sensitivity matrix around the current state. An important effect of this approach is that the residuals $(\mathbf{y}_\mathrm{k} - \mathbf{M}_\mathrm{k}\mathbf{x}_\mathrm{k}^{-})$ become approximately normally distributed, a prerequisite for the Kalman Filter to provide an optimal solution. Finally, temporal correlations in the residuals were accounted for by applying an augmented state red-noise Kalman Filter as described in Brunner et al. (2012).

2.3 Background Treatment

The mole fractions of an inert trace gas at any given point in the atmosphere may be considered to be composed of a smoothly varying, large-scale background plus a more rapidly varying component containing the imprint of recent sources and sinks. Since the LPDM simulations only account for the contribution from recent emissions (the time period covered by the backward simulations), the background has to be treated separately. All inversion systems estimated a prior background, and three of the four systems optimized the background along with the emissions, but the details of this optimization differed.

For the prior background mole fractions, NILU used the method described in Thompson and Stohl (2014). In brief, this involved the following three steps: 1) selecting observations defined to be representative of the background, i.e., lower quartile of values in a shifting time window of 60 days (30 days for SF_6), 2) calculating the contribution to these observations from prior emissions within the domain and subtracting these, and 3) interpolating the background mole fractions to the observation time step.

EMPA2 applied the Robust Estimation of Baseline Signal (REBS) method (Ruckstuhl et al., 2012), which iteratively fits a non-parametric local regression curve to the observations, successively excluding points outside a certain range around the baseline curve. REBS was applied separately to individual observations from each site using asymmetric robustness weights with a tuning factor of b =2.5, a temporal window width of 60 days and a maximum of 10 iterations. An estimate of the baseline uncertainty is given by REBS as a constant value for the whole time series.

In the UKMO set up, a total of eleven extra ‘boundary condition’ variables were estimated as part of the inversion. The prior background time-series was calculated using data at Mace Head when well-mixed ‘clean’ air arrived from the North Atlantic Ocean. The eleven variables are multiplication factors to calculate the mole fractions of the background air arriving from eight horizontal (SSE, SSW, WSW,…, ESE) boundaries at 0-6 km, two boundaries (north and south) from 6 to 9 km, and a boundary at 9 km (upper troposphere to stratosphere).

EMPA2 optimized the REBS background levels separately for each measurement site at selected reference points every 14 days. The uncertainty provided by the REBS procedure served as prior uncertainty during the inversion. Background levels in between these reference points were linearly interpolated. NILU did not optimize the background to avoid cross-talk between the optimization of the emissions and the baseline. In the EMPA system, a single element per observation site is added to the state vector to represent the background at time step *k*. This background is then allowed to evolve slowly with time similar to the evolution of the emissions (see Eq. 7). As first guess for the initialization of the assimilation, the 5th percentile of the first 12 days of measurements is used.

2.4 Inversion Grids

In order to limit the dimension of the problem, all four systems feature a reduced resolution grid to represent the emissions in the state vector. EMPA and EMPA2 computed a reduced grid by iteratively aggregating grid cells until the enlarged cell passed a threshold with respect to its annual mean total surface sensitivity. The result of this procedure is illustrated in Figure 1, which also presents the position of the three measurement sites and the common domain chosen for the inversion.

NILU employed a reduced grid based on the emission sensitivity with a maximum resolution of 1°x1° over land (effectively most of Europe is resolved at 1°x1° and larger grid cells are only found in Eastern Europe), and a resolution of 4°x4° over sea. UKMO used a grid that follows the outlines of countries or groups of countries of interest, which ensures that parts of different countries are prevented from being aggregated into the same coarse grid. Within country, grid cells can be split further depending on the sensitivity of the measurements to emissions from such areas.

2.5 Experiments

All experiments and required outputs were described in a detailed modelling protocol available to the participants. Three main experiments (M1-M3) were defined to estimate the emissions of HFC-125, HFC-134a, and SF_6, respectively. For HFC-125, several additional experiments were defined to test the sensitivity to changing prior uncertainty, background treatment, data selection, and uniform versus spatially resolved prior emissions. Most of these sensitivity tests were limited to a single inversion system. A summary of the main and sensitivity experiments is presented in Table 2. All experiments were performed for a single year (2011) and the main scope was the estimation of annual mean emissions.

To make the results as comparable as possible, all inversion systems used the same observation data (including uncertainties) and prior emissions, and the backward transport simulations were started from the same horizontal coordinates. Since the comparatively coarse topography in the transport models significantly underestimates the true altitude of the two mountains sites, particles were released at 3000 m amsl at Jungfraujoch and at 2000 m amsl at Monte Cimone, thus a few hundred meters below the true station height but still well above the model topography. Previous analyses of FLEXPART simulations indicated that 3000 m amsl is an optimal release height for Jungfraujoch at the given model resolution of 0.2° x 0.2° (Brunner et al., 2012). However, for the NAME model it turned out that a release height of 3000 m amsl. overestimates the sensitivity to regions surrounding Jungfraujoch, especially France. For NAME a significantly higher release height of 2000 m above model ground (which corresponds to 3906 m amsl) was selected to provide footprint sensitivities comparable to those of FLEXPART.

In order to preserve the characteristics of the individual inversion systems as used in previous studies, no further common settings were specified. In particular, the groups were free to choose the inversion grid, the prior uncertainties (except for experiment FLAT) and error correlation structures (see Table 1). Model outputs defined by the protocol included simulated time series at the measurement sites, gridded emission fields, and estimates of country-aggregated emissions. These outputs form the basis of the results presented in the following.

3. Results and Discussion

3.1 Simulated Time Series

Simulated prior and posterior time series at all three measurement sites are shown in Figure 2 and 3 for HFC-125 mole fractions for experiment M1 (definition see Table 2). Corresponding figures for M2 (HFC-134a) and M3 (SF_6) are presented in the supplementary material.

The simulations successfully reproduce much of the observed variability, indicating that the underlying variations in meteorology and atmospheric transport are well represented by the models. The variance explained by the prior time series ranges between 30% and 80% depending on the site (lowest at Monte Cimone, highest at Mace Head) and the LPDM, and is further increased in the posterior time series. The alternation between clean Atlantic air and advection of polluted air masses from UK and the European continent observed at Mace Head is very well matched by all models. The largest difference between the models is the representation of background concentrations, with NILU being lower than the other models towards the end of the one-year period at Mace Head. The two mountain sites Jungfraujoch and Monte Cimone are more frequently perturbed by polluted air masses and the background level is less clearly defined. As a consequence, the scatter between the background levels is rather large with UKMO tending to be at the lower and EMPA at the upper end of the estimates. Note, however, that EMPA does not have a prior background in the same way as the other models since its background is constructed directly during the assimilation process. The prior mole fractions shown in Figure 2, therefore, have been added to the posterior background in the case of EMPA.

Although many of the peaks observed at the two mountain sites are well captured, reproducing the observations is more challenging at these sites compared to Mace Head. At all three sites, the performance of the posterior simulations is clearly improved and the spread between model simulated peaks and background levels is reduced.

The overall model performances in experiments M1-M3 are summarized in Figure 4 in the form of Taylor diagrams. For HFC-125, the diagrams confirm the qualitative picture presented above: Mace Head is simulated best with posterior correlations between 0.8 and 0.92, compared to values in the range of 0.6 to 0.82 at the mountain sites. The posterior scores are closer to each other than the prior scores. In particular, the score of the NAME-based system UKMO is moving closer to the three FLEXPART-based systems EMPA, EMPA2, and NILU. For HFC-134a, the posterior performances are similar as for HFC-125 except for Monte Cimone where all models have difficulties in reproducing the observations. While the prior simulations of HFC-125 showed too little variance at Jungfraujoch and Mace Head suggesting that emissions in the surroundings of these sites were underestimated, the prior simulations of HFC-134a tended to be too high. Observations of SF_6 were only available from Jungfraujoch and Mace Head. SF_6 is very well simulated at these sites such that the improvement from prior to posterior is relatively small.

Overall, the FLEXPART-based systems performed somewhat better than the UKMO system. This is especially true for Jungfraujoch whereas at Mace Head the differences were minor. The reasons for this are unclear: Differences in the dispersion model, the underlying meteorological model, and/or model setup (e.g. particle release height) are all potential candidates for further study.

3.2 Gridded emissions

Gridded prior emissions are exemplarily presented in Figure 5 for HFC-134a (experiment M2). Although based on exactly the same EDGAR v4.2 inventory data, which 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. The UKMO grid, for example, is rather coarse and follows the country outlines as closely as possible given the resolution of EDGAR v4.2. The grids of NAME, EMPA and EMPA2 have higher resolution (up to 0.1°, see Table 1) near the observation sites and lower resolution further away. NILU has a nearly constant resolution over land and reduced resolution over the sea. These different grids combined with different a priori uncertainties and correlation length scales will influence the inversion results as they offer different flexibility to optimize the emissions. Further insights into these sensitivities will be presented in Sect. 3.4 (country aggregated emissions).

The emission updates, i.e., the posterior minus prior emissions are shown in Figures 6–8 for experiments M1 to M3. For HFC-125, the posterior differences share a number of similarities between the models such as positive values over the Iberian Peninsula, mid and southern Italy, western France, south-western UK, and negative values over northern Italy and northern/north-eastern UK. Overall, EMPA and EMPA2 are quite similar except for opposing patterns over the Benelux countries and south-eastern UK. NILU estimates much larger enhancements over Spain than the other models. It also finds significant enhancements in a band extending from Germany towards the Baltic countries, where the other models find either small (UKMO) or even negative increments (EMPA, EMPA2). These rather large differences are somewhat surprising considering the fact that the posterior time series simulated by the models are of similar quality (Figure 3). A notable difference between the models is the consistently lower background in the NILU system at Mace Head between October and December, probably because it does not optimize the background in the inversion. However, the sensitivity test NOBLOPT

(Table 2, results in Sect. 3.4), where EMPA2 repeated the experiment without background adjustment, still showed large differences from NILU in this period, suggesting that they were already present in the prior background. In the case of no background optimisation, emissions estimated by EMPA2 were generally higher in most of the domain (total of 1.1 Gg/yr higher) as compared with the reference run M1. Differences were especially large for the Iberian Peninsula and Italy, but not towards north-eastern Europe as in NILU.

A similar picture emerges for HFC-134a (Figure 7). The models estimate reductions with respect to the prior emissions over eastern and northern UK and northern Italy. All models find enhanced posterior emissions over Spain and Portugal with NILU estimating again the largest changes, similar to HFC-125. For Germany, there is little consistency between the models. While NILU and EMPA show reductions over the western and increases over the eastern parts of the country, EMPA2 estimates a uniform reduction and UKMO finds decreases in the northern and increases in the southern parts. A unique feature of NILU is again a band of positive changes extending from Germany to the Baltic countries. UKMO simulates a pronounced dipole pattern in the area of Paris. Such dipole patterns occur more easily when spatial correlations in the prior uncertainties are not considered.

For SF_6, all models consistently simulate lower posterior than prior emissions over Germany, the country with the largest emissions of SF_6 in Europe. Except for UKMO, the models consistently find increased emissions in Italy and the western parts of France. Similar to HFC-125 and HFC-134a but different from the other systems, NILU simulates strong enhancements for the Iberian Peninsula. Most models find a local reduction around Jungfraujoch, especially UKMO.

3.3 Uncertainty reductions

A useful diagnostic of the model results is the uncertainty reduction as it illustrates the influence of the measurements on the posterior fields. 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.

Figure 9 presents the absolute prior uncertainties chosen in the four systems for the example of HFC-134a. Corresponding figures for HFC-125 and SF_6 are provided in the supplement. EMPA and EMPA2 specified the uncertainties relative to the prior emissions. As a result, the distribution closely follows the pattern of prior emissions. This is also true for UKMO although uncertainties in grid cells with very low emissions were set to a minimum value. Overall, much lower prior uncertainties were specified in EMPA and EMPA2 compared to NILU and UKMO. In EMPA, the relative uncertainties were set to a range of about 70% for the largest and 100% for the smallest grid cells, accounting for the assumed uncertainty correlation length of 500 km. In EMPA2, the uncertainties were set uniformly to 137%, but to prevent negative emissions, these uncertainties had to be reduced iteratively in some grid cells. UKMO assumed a 200% uncertainty in the prior emissions plus a minimum value. In NILU the uncertainties for each grid cell were set to 100% of the largest emission out of itself and the 8 neighbouring grid cells and in addition, a minimum uncertainty was specified. This was done to allow a higher degree of freedom in adjusting the spatial pattern of emissions.

Together with the different spatial uncertainty correlations, these differences have a marked effect on the resulting uncertainty reductions. Figure 10 shows the reductions achieved for HFC-134a. Uncertainty reductions are largest and rather uniform for NILU due to the large prior uncertainties and prior error correlations with a length scale of 200 km over land. Almost no reductions are found over sea due to very low prior uncertainties. Uncertainty reductions are more scattered in EMPA2 due to the absence of spatial correlations in the prior error covariance matrix. The pattern reflects a combination of the influence of the measurements and magnitude of the prior fluxes. Largest reductions tend to occur in grid cells with large prior emissions. Due to the growing cell sizes with increasing distance from the measurements, error reductions do not fall off as clearly with distance from the sites as in the NILU system.

Uncertainty reductions are only moderate in UKMO despite rather large prior uncertainties. This is likely due to an eight times smaller number of observations assimilated (one morning or afternoon value instead of eight 3-hourly values per day) compared to EMPA and EMPA2 and larger assumed data-mismatch uncertainties, especially compared to NILU. The data-mismatch uncertainties adopted for Mace Head, for example, correspond to average HFC-134a mole fraction uncertainties of 1.9 ppt for EMPA and EMPA2, 1.2 ppt for NILU, and 3.4 ppt for UKMO, respectively. At Jungfraujoch, the uncertainty specified in UKMO was about 5 times larger than in the other models, reflecting the high uncertainty in simulated transport assumed for this site. Note that in all inversion systems the data-mismatch uncertainty is much larger than the stated measurement precision and is thus dominated by representation and transport model uncertainties.

Due to the optimization of the logarithm of emissions, the EMPA system reduces relative rather than absolute uncertainties. The uncertainty reduction is, therefore, presented in terms of reduction of relative uncertainties. The uncertainty reductions are typically between 40% and 70%. Similar to EMPA2, the uncertainty reductions do not fall off strongly with distance from the sites due to the irregular grid. Unlike EMPA2, however, the pattern is much more uniform due to the consideration of spatial error correlations. Minor maxima coincide with grid cells with large prior emissions.

3.4 Country aggregated emissions

An important question in the context of international treaties such as the recent Paris Agreement is, how suitable is the current observation network to constrain emissions at the country level? For this purpose, the gridded emission fields were aggregated to individual countries or groups of countries. Due to the relatively coarse grids, this aggregation can be a significant source of error. Emissions from grid cells covering two or more countries need to be properly assigned to the individual countries. This was done either by weighting according to the fractional area covered by each country (EMPA, NILU), or by weighting according to the relative share of the population in the overlapping cell using high-resolution population density data (EMPA2). UKMO circumvented the problem by specifying a grid following the country borders.

Another critical question is whether emissions from grid cells covering both land and sea should be fully assigned to the land areas or whether only the fraction covered by land should be considered. This is particularly relevant for countries like Italy with long coastlines and for inversion grids with large cells. In all models it was assumed that emissions from grid cells partially overlapping sea areas are fully assigned to the adjacent land areas assuming that emissions over sea are negligible. UKMO explicitly extended the country masks to include offshore sea areas.

Figure 11 presents the prior emissions of HFC-125 estimated by the four model systems. Differences between these estimates reflect the uncertainty introduced by the different grids and country attribution strategies. These differences are typically in the range of 1% to 6% of the country emissions but occasionally can be larger. For Denmark, for example, the values vary between a minimum of 32 Mg/yr (EMPA) and 120 Mg/yr (UKMO). The low value estimated by EMPA is largely attributable to the area of Copenhagen being part of a large grid cell also covering large parts of southern Sweden resulting in a significant misattribution of emissions from Denmark to Sweden. As a consequence, emissions from SW+FI+BALT are relatively high in this model. Estimates of EMPA2 and UKMO are generally very close to each other suggesting that the usage of high-resolution population density data for redistributing sub-grid cell emissions is nearly equivalent to using a grid following the country outlines.

The corresponding posterior estimates for HFC-125 are shown in Figure 12. Here, the differences between the models are much larger. EMPA and NILU have larger adjustments with respect to the prior than the other two models; integrated over all countries their emissions are about 50% higher. The standard deviation between the four model estimates for the domain total is 26%. NILU estimates particularly large enhancements for Germany, the Iberian countries ES+PT, the Nordic countries SW+FI+BALT, and the eastern European countries PO+CZ+SV, consistent with the spatial pattern in Figure 6. EMPA, conversely, estimates only small changes for Germany, similarly large enhancements for ES+PO, and uniquely large enhancements for Italy and the Benelux countries BE+NL+LU. The stronger adjustments in EMPA and NILU are likely related to the spatial error correlations considered in these systems but also to other factors (see Sect. 3.5).

Rather than considering the models individually, they may also be treated as an ensemble of estimates that can be compared to the bottom-up emissions officially reported to UNFCCC. A summary of this comparison for the experiments M1-M3 as well as the sensitivity experiment FLAT (discussed in Sect. 3.5) is presented in Figure 13. Shown are median values for the prior and posterior model estimates as well as the range between minimum and maximum. For HFC-125 (panel a) there is a rather high consistency between the top-down estimates and the UNFCCC values for many countries including FR, IT, UK, and Benelux. Marked differences with all models being either higher or lower than UNFCCC are found for DE (model median is 2.4x higher than UNFCCC), ES+PT (4.9x higher), IR (9.5x higher), SW+FI+BALT (2x higher), PO+CZ+SV (2.8x smaller), and CH (2x smaller). It should be noted that the prior emissions based on the EDGAR v4.2 2008 inventory for HFC-125 are significantly different from the UNFCCC 2011 emissions officially reported by the countries (grey bars). This is especially true for the countries DE and PO+CZ+SV, where the posterior model estimates are closer to the EDGAR prior. The estimated significant underestimation of the HFC-125 emissions reported to UNFCCC by Ireland and Spain+Portugal, that was consistently found across all model systems, has also been reported previously by Brunner et al. (2012). Summed over all countries, the model median estimate is 24% higher than the UNFCCC total. For some countries, our results can also be compared with those by Lunt et al. (2015), which covered a similar period (2010-2012) and also used EDGAR as prior (see their Table S3). For example, they also found higher than UNFCCC emissions from Germany though not as large as EDGAR. For France their posterior remained close to EDGAR and was lower than UNFCCC. Emissions from UK and Italy were significantly increased which is in contrast to our results.

For HFC-134a, the model estimates are generally more consistent with UNFCCC than for HFC-125 (Figure 13c). In strong contrast to HFC-125, this is also true for Ireland and Spain+Portugal. The high consistency also applies to the domain total, which is only 11% lower than the total reported to UNFCCC. For SW+FI+BALT and PO+CZ+SV there are similar discrepancies as for HFC-125. Again, this is at least partly caused by the large differences between the prior and UNFCCC emissions and the large influence of the prior on the final model estimates. The model estimates are consistently lower than the UNFCCC values for UK by about a factor of two, which contributes strongly to the 11% difference for the domain total. An overestimation of the HFC-134a emissions reported by UK has also been found previously by Lunt et al. (2015) and Say et al. (2016) and is in part due to the use of an assumed high loss rate of HFC-134a from car air conditioning systems in the UK. For Italy, the model estimates are consistently higher than the UNFCCC values by 40% on average. Note, however, that the results for Italy are strongly influenced by the measurements at Monte Cimone where the models had difficulties in reproducing the HFC-134a measurements. Lunt et al. (2015) found an even stronger increase over Italy (factor 2.4), whereas they obtained relatively consistent (compared with UNFCCC) estimates for Germany and reductions by ~25 % in France, in fair agreement with our results.

For emissions of SF_6 the attribution to the different countries is very different from HFC-125 and HFC-134 (Fig. 13d). Consistent with the bottom-up estimates reported to UNFCCC, the models identify Germany as the highest national emitter in Europe. The model median is highly consistent with UNFCCC but almost a factor 2 lower than the EDGAR v4.2 prior. For almost all other countries, however, the model estimates are closer to EDGAR v4.2 than to UNFCCC. For Italy, France, and Spain+Portugal, for example, the model medians are a factor 2-3 higher than the UNFCCC values but very close to EDGAR v4.2. Summed over all countries, the models are 47% higher than UNFCCC. SF_6 emissions have also been estimated by Ganesan et al. (2014) for the year 2012 based on a slightly modified EDGAR4.2 prior. Their estimates for Germany (348 Mg/yr) were much higher than ours (137 Mg/yr), but also their prior was much higher (650 Mg/yr compared to 254 Mg/yr). We note that our prior (obtained as a sum over all grid cells covering Germany) is consistent with the country table provided by the EDGAR inventory.

3.5 Sensitivity to different model assumptions

A set of additional HFC-125 experiments was conducted by a subset of models to analyse the sensitivity to different assumptions and identify possible reasons for the model-to-model differences (Table 2). A first test conducted by all models was an experiment for HFC-125 similar to M1 but using a flat, non-informative prior (FLAT), which had one emission value over land and one over ocean, to test the ability of the models to reconstruct the spatial distribution of emissions with no corresponding prior information. In this experiment, the uncertainty for the domain total emissions was set to 100%. Other experiments included tests with doubled (U200%) and halved (U50%) prior uncertainties conducted by NILU and UKMO, two tests with no optimization of the baseline conducted by EMPA2, the first one using EMPA2's baseline (NOBLOPT) and the second one using NILU's somewhat lower baseline (NILUBL), and tests with daily mean (DMEAN) and one single observation per day (ONEOBS) instead of 3-hourly observations conducted by EMPA to mimic the sampling of NILU and UKMO.

The estimates with a flat prior (Figure 13, panel b) are similar to those with the spatially explicit prior (panel a) for most countries well covered by the footprint of the three measurement stations, notably for DE, IT, FR, UK, and IR, suggesting that the model ensemble provides a robust estimate for these countries that is mainly informed by the measurements rather than the prior. This is less true for the individual models as shown in Table 3, which summarizes the results of all experiments for the largest well-covered countries in the domain. For countries in the east and northeast of the domain (SW+FI+BALT, NO, PO+CZ+SV), which are poorly 'seen' by the three sites, the median posterior remains close to the prior, and the posterior differences between experiments FLAT and M1 resemble the prior differences. For ES+PT both priors are too low, but starting from a higher prior (experiment FLAT) results in an even higher posterior, especially in EMPA2 and UKMO.

Comparing the range of individual model estimates (Table 3 and uncertainty bars in Figure 13) suggests that model-to-model differences were of similar magnitude in experiments FLAT and M1 despite a more uniform setup in FLAT with an agreed total uncertainty. The differences thus appear to be mainly caused by the many other choices such as spatial correlations of the prior, grid structure, background treatment, magnitude and correlation structure of the observation uncertainties, and transport model.

Some further insight is provided by the other sensitivity simulations: Decreasing or increasing the prior uncertainties by a factor of two relative to M1 changed the country estimates by only about 10% or less (Table 3). An exception is ES+PT where the results depended strongly on the prior uncertainty, which is a clear indication that the emissions from the Iberian countries are not well constrained by the current observation network. Switching off the baseline optimization in EMPA2 to mimic the setup of NILU increased the emissions in all countries between +6% (DE) and up to +19% (FR, ES+PT). This indicates that with optimization the baseline in EMPA2 tended to be corrected upward and that without optimization this had to be compensated by higher emissions. In a further sensitivity experiment conducted by EMPA2 with no optimization, EMPA2's baseline was replaced by NILU's baseline, which tends to be lower due to the subtraction of simulated mole fractions from the background values (see Sect. 2.3). This further increases the emissions in almost all countries, most strongly in France (+117% with respect to experiment M1) followed by Span+Portugal (+55%) and Italy (+35%), whereas in Germany and UK the changes are small. Despite using the same baseline, the spatial pattern of emission adjustments does not bring EMPA2 much closer to NILU (not shown). In particular, the large positive changes over Germany are not reproduced and those over Italy and France become more strongly positive compared to NILU. This suggests that the baseline selection is not the only factor explaining the differences between EMPA2 and NILU, but that the amplitude and correlation structure of the prior uncertainties as well as the grid geometry are also contributing.

Finally, the influence of different sampling and averaging of the observations was tested with the EMPA system in experiments DMEAN and ONEOBS to mimic the sampling of NILU and UKMO, respectively. Note that for experiment DMEAN the model-data mismatch uncertainty was reduced to respect the requirement of a χ^2 value close to the number of observations (Brunner et al., 2006). The results for DE and IT changed only little but they changed substantially for FR, UK and ES+PT. With daily averaged instead of 3-hourly observations the estimate for FR increased by 17%, and with one observation per day decreased by 22%, the latter being closer to the prior. For the UK, however, the opposite effect is seen, with daily means reducing (-13%) and one-observation-per-day increasing (+31%) the estimate relative to M1. The results for the UK are dominated by observations from the station Mace Head. At this site, the mean diurnal cycle of the differences between FLEXPART simulated and observed concentrations exhibits negative differences (-0.07 ppt) in the afternoon but positive differences (0.02-0.05 ppt) during the rest of the day. When using only afternoon observations as in experiment ONEOBS and as used by UKMO, the EMPA system thus requires higher emissions to compensate for the negative bias compared to when all data are used. Both experiments suggest a considerable impact of the choice of observations, which is in contrast to previous findings of Brunner et al. (2012), who made a similar sensitivity experiment and found only a relatively small influence. Except for the UK, the estimates of experiment DMEAN were always higher than those of experiment ONEOBS, consistent with NILU being generally higher than UKMO. Some of the differences between the model results are thus likely attributable to the specific selection and aggregation of the observation data.

4 Conclusions

For the first time, four independent regional inversion systems for synthetic greenhouse gas emissions have been applied in well-controlled model experiments to compare the systems and to analyse the performance of the ensemble. Emissions of the two most important halocarbons in terms of (CO_2-eq.) greenhouse gas emissions in Europe, HFC-125 and HFC-134a, as well as SF_6 were estimated for the year 2011. The four model systems, referred to as EMPA, EMPA2, NILU, and UKMO, differed in terms of Lagrangian transport model (3 x FLEXPART with ECMWF IFS meteorology, 1 x NAME with UKMO meteorology) and inversion method (3 x Bayesian inversion, 1 x extended Kalman Filter). The inversion systems used the same observation time series and a priori emission fields but differed in a number of other aspects such as the amplitude and correlation structure of the prior and observation uncertainty covariance matrices, the treatment of background mole fractions, the inversion grid and resolution, and the averaging or subsampling of observations, in order to preserve the characteristics of the individual approaches as used in previous studies as much as possible.

All systems were able to reproduce the measurement time series well to very well. Pearson's correlation coefficients for the prior simulations were typically in the range 0.6-0.7 at Jungfraujoch, 0.8-0.9 at Mace Head, and 0.5-0.7 at Monte Cimone. Correlation coefficients for the posterior time series were about 0.05 to 0.1 better and bias-corrected RMSE were typically reduced by 10 to 40% with the exception of HFC-134a at Monte Cimone, where the reduction was only between 2 and 5% in all systems. The transport model NAME was less successful than FLEXPART in reproducing the measurements at the two mountain sites JFJ and CMN but showed comparable performance at MHD.

The comparison of gridded emissions was complicated by the large differences in resolution and structure of the inversion grids: the number of grid elements optimized varied between 150 in the UKMO, 522 in EMPA2, 1083 in EMPA and 1140 in the NILU system. UKMO, EMPA and EMPA2 had a high grid resolution near the measurement sites and lower resolution at larger distance where the measurements were less sensitive, especially over eastern and south-eastern Europe and Scandinavia. The UKMO grid followed the country borders to simplify emission attribution to individual countries.

For HFC-125, all inversion systems estimated higher posterior emissions compared to the EDGAR v4.2 prior for the Iberian Peninsula and most of Italy except northern Italy. The models also tended towards higher posterior emissions over Ireland and southwestern UK but lower emissions over the eastern and northern parts of UK. A unique feature of the NILU system was a band of positive posterior – prior differences extending from Germany towards the Baltic countries. For HFC-134a, the patterns of changes were similar but showed more negative posterior – prior differences (e.g., over the Benelux countries and the UK). For SF_6, all models simulated the highest emissions over Germany though much reduced with respect to the EDGAR v4.2 prior. In contrast to Germany, SF_6 emissions for Italy and France were higher than the prior.

Overall, NILU and EMPA tended to retrieve higher emissions than UKMO and EMPA2. For all three gases, NILU had the highest total domain emissions and EMPA2 the lowest. These results are related to two main factors: First, EMPA and NILU were the only systems considering spatial correlations in the prior resulting in a smaller number of degrees of freedom and a correspondingly stronger influence of the observations on the posterior emissions. Second, NILU was the only system not applying a correction to the background in order to avoid cross-talk between the optimization of the emissions and the background. A sensitivity experiment for HFC-125 with no background adjustment conducted by EMPA2 indeed resulted in higher emissions though not reaching the levels of NILU.

The patterns of uncertainty reductions differed strongly: NILU and EMPA had rather smooth reductions whereas the patterns of EMPA2 and UKMO were more scattered due to the absence of spatial correlations in the prior uncertainties. NILU assumed large and rather uniform (absolute) prior uncertainties and, as a result, found the largest uncertainty reductions. UKMO also had large prior uncertainties but much smaller reductions due to their assumption of large observation uncertainties.

Gridded emissions were aggregated to individual countries to analyse the consistency between the models and to compare the results against country totals officially reported to the UNFCCC (reported in 2013 for the year 2011) and to the EDGAR v4.2 prior (representing 2008). The rather coarse inversion grids were a non-negligible source of uncertainty (typically between 1 and 6%) when aggregating the emissions to individual countries. The overall magnitude of the emissions and the attribution to different countries such as the dominant role of Germany for SF_6 emissions were quite consistent with the UNFCCC estimates. However, the estimates of the individual models varied considerably. 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. The individual models often did not overlap within the range of the combined uncertainties suggesting that the analytical uncertainties are a poor representation of the true uncertainties, which are rather dominated by parametric and structural uncertainties.

The ensemble median agreed very well with the UNFCCC estimates for HFC-134a for most countries, better than any single model. As also found in previous studies, emissions of HFC-134a reported to UNFCCC by the UK appear to be about a factor two too high. A similar conclusion may be drawn for the group Poland+Czech Republic+Slovakia though with less confidence due to the limited coverage of these countries by the current observation network. In terms of HFC-125 emissions the largest discrepancies from UNFCCC values were found for Spain+Portugal and for Ireland, with model medians 4.9 times and 9.5 times higher, respectively. Interestingly, for the same countries the model estimates for HFC-134a were highly consistent with the reported values, providing further evidence that the reported HFC-125 emissions are too low.

Consistent with the UNFCCC reports, the models identified Germany as the highest national emitter of SF_6 in Europe. The model estimates for Germany agreed well with the UNFCCC numbers but were a factor 2 to 3 higher for Italy, France and Spain+Portugal.

The current network of three routine monitoring sites for synthetic greenhouse gases in Europe is only able to constrain the broad spatial patterns of their emissions, such as the concentration of SF_6 emissions on Germany as opposed to the more uniform distribution of emissions of HFC-125 and HFC-134a. 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, occasionally as large as the estimated emissions, whereas the median appears to have significant skill as judged from the comparison with reported HFC-134a emissions, which are considered to be relatively well known. The sensitivity experiments were not sufficient to fully disclose the origin of the model-to-model differences, but factors such as subsampling of observations, background treatment, and magnitude and correlation structure of the prior uncertainties were identified as playing an important role. Further work will be needed, for example by testing the model's internal consistency using a χ^2 test, and by separating model transport from other uncertainties, to build trust in the inverse modelling systems.

Acknowledgements

This study was funded by the European Commission's Seventh Framework Programme project InGOS (grant agreement no. 284274). Measurements at Jungfraujoch are supported by the Swiss Federal Office for the Environment (FOEN) through the project HALCLIM and by the International Foundation High Altitude Research Stations Jungfraujoch and Gornergrat (HFSJG). Measurements at Mace Head are supported by the Department of Energy & Climate Change (DECC, UK) (contract GA0201 with the University of Bristol). The O. Vittori station Monte Cimone is supported by the National Research Council of Italy.

References

Baker, D.F., Law, R.M., Gurney, K.R., Rayner, P., Peylin, P., et al. (2006): TransCom 3 inversion intercomparison: Impact of transport model errors on the interannual variability of regional CO2 fluxes, 1988–2003. Glob. Biogeochem. Cycle 20, 1-17,doi: 10.1029/2004GB002439.

Berchet, A., Pison, I., Chevallier, F., Bousquet, P., Bonne, J.L., et al. (2015): Objectified quantification of uncertainties in Bayesian atmospheric inversions. Geosci. Model Dev. 8, 1525-1546,doi: 10.5194/gmd-8-1525-2015.

Berchet, A., Pison, I., Chevallier, F., Bousquet, P., Conil, S., et al. (2013): Towards better error statistics for atmospheric inversions of methane surface fluxes. Atmos. Chem. Phys. 13, 7115-7132,doi: 10.5194/acp-13-7115-2013.

Bergamaschi, P., Corazza, M., Karstens, U., Athanassiadou, M., Thompson, R.L., et al. (2015): Top-down estimates of European CH4 and N2O emissions based on four different inverse models. Atmos. Chem. Phys. 15, 715-736,doi: 10.5194/acp-15-715-2015.

Bergamaschi, P., Krol, M., Meirink, J.F., Dentener, F., Segers, A., et al. (2010): Inverse modeling of European CH4 emissions 2001–2006. J. Geophys. Res. 115, 1-18,doi: 10.1029/2010JD014180.

Brunner, D., Henne, S., Keller, C.A., Reimann, S., Vollmer, M.K., et al. (2012): An extended Kalman-filter for regional scale inverse emission estimation. Atmos. Chem. Phys. 12, 3455-3478,doi: DOI 10.5194/acp-12-3455-2012.

Brunner, D., Staehelin, J., Künsch, H.R., Bodeker, G.E. (2006): A Kalman filter reconstruction of the vertical ozone distribution in an equivalent latitude - potential temperature framework from TOMS/GOME/SBUV total ozone observations. J. Geophys. Res. 11, D12308,doi: 10.1029/2005JD006279.

Cunnold, D.M., Fraser, P.J., Weiss, R.F., Prinn, R.G., Simmonds, P.G., et al. (1994): Global trends and annual releases of CCl3F and CCl2F2 estimated from ALE/GAGE and other measurements from July 1978 to June 1991. J. Geophys. Res. 99, 1107-1126,doi: 10.1029/93JD02715.

Derwent, R.G., Simmonds, P.G., Greally, B.R., O'doherty, S., McCulloch, A., et al. (2007): The phase-in and phase-out of European emissions of HCFC-141b and HCFC-142b under the Montreal Protocol: Evidence from observations at Mace Head, Ireland and Jungfraujoch, Switzerland from 1994 to 2004. Atmos. Environ. 41, 757-767,doi: 10.1016/j.atmosenv.2006.09.009.

Ganesan, A.L., Rigby, M., Zammit-Mangion, A., Manning, A.J., Prinn, R.G., et al. (2014): Characterization of uncertainties in atmospheric trace gas inversions using hierarchical Bayesian methods. Atmos. Chem. Phys. 14, 3855-3864,doi: 10.5194/acp-14-3855-2014.

Graziosi, F., Arduini, J., Furlani, F., Giostra, U., Kuijpers, L.J.M., et al. (2015): European emissions of HCFC-22 based on eleven years of high frequency atmospheric measurements and a Bayesian inversion method. Atmos. Environ. 112, 196-207,doi: http://dx.doi.org/10.1016/j.atmosenv.2015.04.042.

Henne, S., Brunner, D., Oney, B., Leuenberger, M., Eugster, W., et al. (2016): Validation of the Swiss methane emission inventory by atmospheric observations and inverse modelling. Atmos. Chem. Phys. 16, 3683-3710,doi: 10.5194/acp-16-3683-2016.

Hurwitz, M.M., Fleming, E.L., Newman, P.A., Li, F., Mlawer, E., et al. (2015): Ozone depletion by hydrofluorocarbons. Geophysical Research Letters 42, 8686-8692,doi: 10.1002/2015GL065856.

Karstens, U., Schwingshackl, C., Schmithüsen, D., Levin, I. (2015): A process-based 222radon flux map for Europe and its comparison to long-term observations. Atmos. Chem. Phys. 15, 12845-12865,doi: 10.5194/acp-15-12845-2015.

Keller, C.A., Brunner, D., Henne, S., Vollmer, M.K., O'Doherty, S., et al. (2011): Evidence for under-reported western European emissions of the potent greenhouse gas HFC-23. Geophysical Research Letters 38,doi: Artn L15808
Doi 10.1029/2011gl047976.

Keller, C.A., Hill, M., Vollmer, M.K., Henne, S., Brunner, D., et al. (2012): European Emissions of Halogenated Greenhouse Gases Inferred from Atmospheric Measurements. Environ Sci Technol 46, 217-225,doi: Doi 10.1021/Es202453j.

Levin, I., Naegler, T., Heinz, R., Osusko, D., Cuevas, E., et al. (2010): The global SF6 source inferred from long-term high precision atmospheric measurements and its comparison with emission inventories. Atmos. Chem. Phys. 10, 2655-2662,doi: 10.5194/acp-10-2655-2010.

Lin, J.C., Gerbig, C. (2005): Accounting for the effect of transport errors on tracer inversions. Geophys. Res. Lett. 32, L01802,doi: 10.1029/2004GL021127.

Locatelli, R., Bousquet, P., Chevallier, F., Fortems-Cheney, A., Szopa, S., et al. (2013): Impact of transport model errors on the global and regional methane emissions estimated by inverse modelling. Atmos. Chem. Phys. 13, 9917-9937,doi: 10.5194/acp-13-9917-2013.

Lunt, M.F., Rigby, M., Ganesan, A.L., Manning, A.J., Prinn, R.G., et al. (2015): Reconciling reported and unreported HFC emissions with atmospheric observations. Proc. Natl. Acad. Sci. USA 112, 5927-5931,doi: 10.1073/pnas.1420247112.

Maione, M., Giostra, U., Arduini, J., Furlani, F., Graziosi, F., et al. (2013): Ten years of continuous observations of stratospheric ozone depleting gases at Monte Cimone (Italy) — Comments on the effectiveness of the Montreal Protocol from a regional perspective. Sci. Tot. Environ. 445–446, 155-164,doi: 10.1016/j.scitotenv.2012.12.056.

Maione, M., Graziosi, F., Arduini, J., Furlani, F., Giostra, U., et al. (2014): Estimates of European emissions of methyl chloroform using a Bayesian inversion method. Atmos. Chem. Phys. 14, 9755-9770,doi: 10.5194/acp-14-9755-2014.

Manning, A.J. (2011): The challenge of estimating regional trace gas emissions from atmospheric observations. Phil. Trans. R. Soc. A 369, 1943-1954,doi: 10.1098/rsta.2010.0321.

Manning, A.J., O'Doherty, S., Jones, A.R., Simmonds, P.G., Derwent, R.G. (2011): Estimating UK methane and nitrous oxide emissions from 1990 to 2007 using an inversion modeling approach. J. Geophys. Res. 116, D02305,doi: 10.1029/2010JD014763.

Manning, A.J., Ryall, D.B., Derwent, R.G., Simmonds, P.G., O'Doherty, S. (2003): Estimating European emissions of ozone-depleting and greenhouse gases using observations and a modeling back-attribution technique. J. Geophys. Res. 108, 4405,doi: 10.1029/2002JD002312.

Michalak, A.M., Hirsch, A., Bruhwiler, L., Gurney, K.R., Peters, W., et al. (2005): Maximum likelihood estimation of covariance parameters for Bayesian atmospheric trace gas surface flux inversions. J. Geophys. Res. 110, D24107,doi: 10.1029/2005JD005970.

Miller, B.R., Weiss, R.F., Salameh, P.K., Tanhua, T., Greally, B.R., et al. (2008): Medusa: A Sample Preconcentration and GC/MS Detector System for in Situ Measurements of Atmospheric Trace Halocarbons, Hydrocarbons, and Sulfur Compounds. Analytical Chemistry 80, 1536-1545,doi: 10.1021/ac702084k.

Montzka, S.A., McFarland, M., Andersen, S.O., Miller, B.R., Fahey, D.W., et al. (2015): Recent Trends in Global Emissions of Hydrochlorofluorocarbons and Hydrofluorocarbons: Reflecting on the 2007 Adjustments to the Montreal Protocol. J. Phys. Chem. A 119, 4439-4449,doi: 10.1021/jp5097376.

Nisbet, E., Weiss, R. (2010): Top-Down Versus Bottom-Up. Science 328, 1241-1243,doi: 10.1126/science.1189936.

O'Doherty, S., Cunnold, D.M., Miller, B.R., Mühle, J., McCulloch, A., et al. (2009): Global and regional emissions of HFC-125 (CHF2CF3) from in situ and air archive atmospheric observations at AGAGE and SOGE observatories. J. Geophys. Res. 114, D23304,doi: 10.1029/2009JD012184.

Rigby, M., Manning, A.J., Prinn, R.G. (2011): Inversion of long-lived trace gas emissions using combined Eulerian and Lagrangian chemical transport models. Atmos. Chem. Phys. 11, 9887-9898,doi: 10.5194/acp-11-9887-2011.

Rigby, M., Mühle, J., Miller, B.R., Prinn, R.G., Krummel, P.B., et al. (2010): History of atmospheric SF6 from 1973 to 2008. Atmos. Chem. Phys. 10, 10305-10320,doi: 10.5194/acp-10-10305-2010.

Ruckstuhl, A.F., Henne, S., Reimann, S., Steinbacher, M., Vollmer, M.K., et al. (2012): Robust extraction of baseline signal of atmospheric trace species using local regression. Atmos. Meas. Tech. 5, 2613-2624,doi: 10.5194/amt-5-2613-2012.

Ryall, D.B., Maryon, R.H. (1998): Validation of the UK Met. Office's name model against the ETEX dataset. Atmos. Environ. 32, 4265-4276,doi: http://dx.doi.org/10.1016/S1352-2310(98)00177-0.

Saikawa, E., Rigby, M., Prinn, R.G., Montzka, S.A., Miller, B.R., et al. (2012): Global and regional emission estimates for HCFC-22. Atmos. Chem. Phys. 12, 10033-10050,doi: 10.5194/acp-12-10033-2012.

Say, D., Manning, A.J., O'Doherty, S., Rigby, M., Young, D., et al. (2016): Re-Evaluation of the UK's HFC-134a Emissions Inventory Based on Atmospheric Observations. Environ Sci Technol 50, 11129-11136,doi: 10.1021/acs.est.6b03630.

Seibert, P., Frank, A. (2004): Source-receptor matrix calculation with a Lagrangian particle dispersion model in backward mode. Atmos. Chem. Phys. 4, 51-63,doi: 10.5194/acp-4-51-2004.

Simmonds, P.G., Rigby, M., Manning, A.J., Lunt, M.F., O'Doherty, S., et al. (2016): Global and regional emissions estimates of 1,1-difluoroethane (HFC-152a, CH3CHF2) from in situ and air archive observations. Atmos. Chem. Phys. 16, 365-382,doi: 10.5194/acp-16-365-2016.

Stohl, A., Forster, C., Frank, A., Seibert, P., Wotawa, G. (2005): Technical note: The Lagrangian particle dispersion model FLEXPART version 6.2. Atmos. Chem. Phys. 5, 2461-2474,doi: 10.5194/acp-5-2461-2005.

Stohl, A., Seibert, P., Arduini, J., Eckhardt, S., Fraser, P., et al. (2009): An analytical inversion method for determining regional and global emissions of greenhouse gases: Sensitivity studies and application to halocarbons. Atmos. Chem. Phys. 9, 1597-1620,doi: 10.5194/acp-9-1597-2009.

Tarantola, A., 2005. Inverse Problem Theory and Methods for Model Parameter Estimation. Society for Industrial and Applied Mathematics, Philadelphia, USA.

Thacker, W.C. (2007): Data assimilation with inequality constraints. Ocean Modelling 16, 264-276,doi: http://dx.doi.org/10.1016/j.ocemod.2006.11.001.

Thompson, R.L., Stohl, A. (2014): 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.

Van dop, H., Addis, R., Fraser, G., Girardi, F., Graziani, G., et al. (1998): ETEX: A European tracer experiment; observations, dispersion modelling and emergency response. Atmos. Environ. 32, 4089-4094,doi: 10.1016/S1352-2310(98)00248-9.

Velders, G.J.M., Fahey, D.W., Daniel, J.S., McFarland, M., Andersen, S.O. (2009): The large contribution of projected HFC emissions to future climate forcing. Proc. Natl. Acad. Sci. USA 106, 10949-10954,doi: 10.1073/pnas.0902817106.

Villani, M.G., Bergamaschi, P., Krol, M., Meirink, J.F., Dentener, F. (2010): Inverse modeling of European CH4 emissions: sensitivity to the observational network. Atmos. Chem. Phys. 10, 1249-1267,doi: 10.5194/acp-10-1249-2010.

Vollmer, M.K., Miller, B.R., Rigby, M., Reimann, S., Mühle, J., et al. (2011): Atmospheric histories and global emissions of the anthropogenic hydrofluorocarbons HFC-365mfc, HFC-245fa, HFC-227ea, and HFC-236fa. Journal of Geophysical Research: Atmospheres 116, n/a-n/a,doi: 10.1029/2010JD015309.

Xiang, B., Patra, P.K., Montzka, S.A., Miller, S.M., Elkins, J.W., et al. (2014): Global emissions of refrigerants HCFC-22 and HFC-134a: Unforeseen seasonal contributions. Proc. Natl. Acad. Sci. USA 111, 17379-17384,doi: 10.1073/pnas.1417372111.

Table 1: Overview of inversion systems

Characteristic	EMPA	EMPA2	NILU	UKMO
Approach	Extended Kalman Filter (ExKF)	Bayesian	Bayesian	Bayesian
Transport model	FLEXPART	FLEXPART	FLEXPART	NAME
Meteorology	ECMWF analyses 0.2°x0.2°, 3hrly	ECMWF analyses 0.2°x0.2°, 3hrly	ECMWF analyses 0.2°x0.2°, 3hrly	UKMO analyses 0.352° x 0.234°, 3hrly
Computational domain	Nested, global	Nested, global	Nested, global	45°W - 40°E, 25°N - 80°N
Inversion grid	0.1°x0.1°minimum, reduced according to residence time	0.1°x0.1°minimum, reduced according to residence time	1°x1° over land, reduced over ocean and far eastern boundary	0.352° x 0.234° min., reduced according to residence time and within country boundaries
State vector length (e=emiss., b=backg., o=other)	1083*e* + 3*b* + 6*o*	522 e + 84 b (405 e + 56 b for M3)	1140e	150e + 11 b
Assimilation time resolution	3-hourly means	3-hourly means	Daily means	3-hourly means once per day
Spatial correlation of prior	500 km	None	200 km over land 1000 km over sea	None
Backwards mode run time	5 days	5 days	10 days	19 days
Prior background mole factions	None, continuously estimated by ExKF	60-day REBS window, biweekly reference points	See Thompson and Stohl (2014) and description below.	Mace Head baseline for all sites, see Manning et al. (2011)
Temporal correlation of observation error	Red-noise Kalman filter	None	None, assumed negligible for daily means	None, assumed negligible with one value per day
Key references	Brunner et al., 2012	Stohl et al., 2009, Vollmer et al., 2009	Thompson and Stohl, 2014	Manning et al., 2011

Table 2: Main (M1-M3) and sensitivity inversion experiments

ID	Gas	Prior inventory	Description	Groups
M1	HFC-125	EDGARv4.2 2008	Reference inversion for HFC-125 for 2011	All
M2	HFC-134a	EDGARv4.2 2008	Reference inversion for HFC-134a for 2011	All
M3	SF_6	EDGARv4.2 2008	Reference inversion for SF_6 for 2011	All
FLAT	HFC-125	Uniform prior[1]	Spatially uniform prior instead of EDGAR	All
U50%	HFC-125	EDGARv4.2 2008	Prior uncertainty reduced by factor 2	UKMO, NILU
U200%	HFC-125	EDGARv4.2 2008	Prior uncertainty increased by factor 2	UKMO, NILU
NOBLOPT	HFC-125	EDGARv4.2 2008	No baseline optimization	EMPA2
NILUBL	HFC-125	EDGARv4.2 2008	Same baseline as NILU, no optimization	EMPA2
DMEAN	HFC-125	EDGARv4.2 2008	Daily means instead of 3-hourly	EMPA
ONEOBS	HFC-125	EDGARv4.2 2008	One instead of eight observations per day	EMPA

[1] One value over land and one value over sea

Table 3: Emissions of HFC-125 in the main experiment M1 and the different sensitivity experiments for major countries in western Europe. UNFCCC refers to the 2011 emissions according to the country reports submitted to UNFCCC in 2013. EDGAR v4.2 refers to 2008 emissions according to the gridding method applied by EMPA2. Uncertainties are shown as ±1σ estimates**.**

Exp. ID	**Model/Inventory**	**DE**	**IT**	**FR**	**UK**	**ES+PT**
		(Mg yr^{-1})	(Mg yr^{-1})	(Mg yr^{-1})	(Mg yr^{-1})	(Mg yr^{-1})
	UNFCCC 2011	*548*	*1169*	*1234*	*1061*	*390*
	EDGAR v4.2 2008	*1232*	*801*	*1001*	*793*	*491*
M1	EMPA	1094 ± 237	2138 ± 240	1483 ± 180	918 ± 144	2599 ± 353
	EMPA2	721 ± 196	1212 ± 73	787 ± 100	812 ± 64	1076 ± 121
	NILU	2078 ± 22	1039 ± 7	1195 ± 13	758 ± 13	2849 ± 17
	UKMO	1568 ± 327	1021 ± 102	919 ± 123	702 ± 235	1218 ± 136
	Median	*1331*	*1125*	*1057*	*785*	*1909*
	Range (min-max)	*721-2078*	*1021-2138*	*787-1483*	*702-918*	*1076-2849*
FLAT	EMPA	1016 ± 354	1522 ± 285	1929 ± 295	1172 ± 273	2713 ± 537
	EMPA2	772 ± 142	1302 ± 149	1067 ± 134	651 ± 94	1769 ± 245
	NILU	1956 ± 20	736 ± 17	1037 ± 17	535 ± 16	2928 ± 29
	UKMO	1586 ± 946	1115 ± 276	1276 ± 298	737 ± 440	3009 ± 499
	Median	*1301*	*1209*	*1172*	*694*	*2820*
	Range (min-max)	*772-1956*	*736-1522*	*1037-1929*	*535-1172*	*1769-2928*
U50%	NILU	2151 ± 21	1055 ± 6	1292 ± 10	766 ± 10	2372 ± 14
	UKMO	1539 ± 195	910 ± 72	824 ± 98	797 ± 145	899 ± 91
U200%	NILU	1936 ± 21	1033 ± 10	1030 ± 14	746 ± 14	3426 ± 19
	UKMO[1]	1422 ± 545	999 ± 165	1066 ± 164	530 ± 330	1739 ± 208
NOBLOPT	EMPA2	770 ± 196	1330 ± 71	937 ± 98	926 ± 64	1284 ± 118
NILUBL	EMPA2	785 ± 181	1643 ± 71	1709 ± 83	837 ± 49	1673 ± 114
DMEAN	EMPA	1123 ± 471	2192 ± 500	1739 ± 399	797 ± 271	2582 ± 780
ONEOBS	EMPA	1068 ± 491	2015 ± 559	1138 ± 337	1209 ± 460	1655 ± 604
	Median	*1488*	*1055*	*1066*	*797*	*1740*
	Range (min-max)	*770-2151*	*910-2192*	*824-1739*	*530-1209*	*899-3426*

[1] Uncertainty increased by 250% rather than 200%

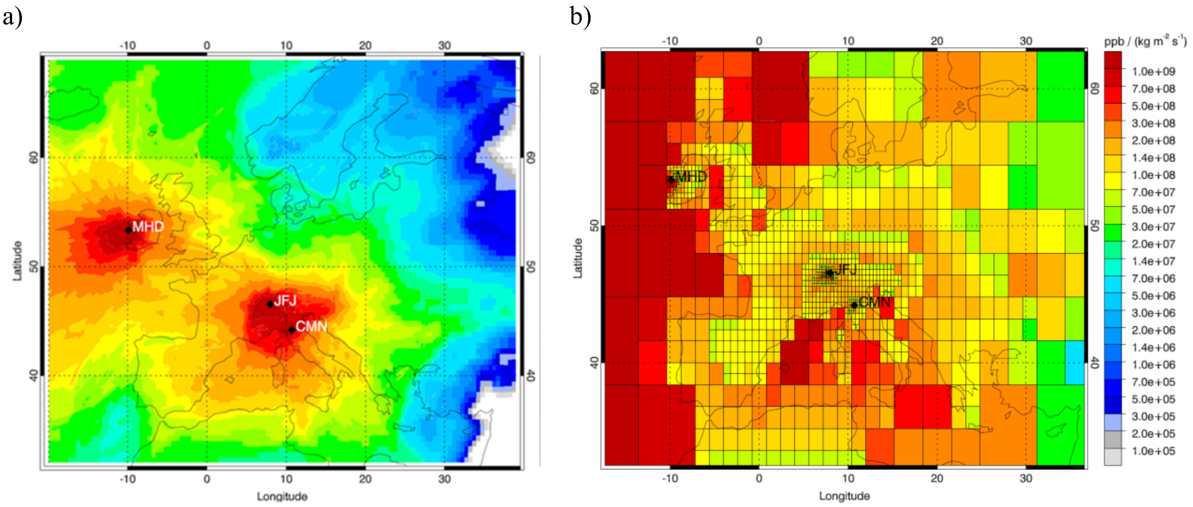


Figure 1: Annual mean surface sensitivity in units of [ppb/(kg m^{-2} s^{-1})] for (a) the original 0.1°x0.1°grid and (b) for the reduced grid of the FLEXPART-based model system EMPA.

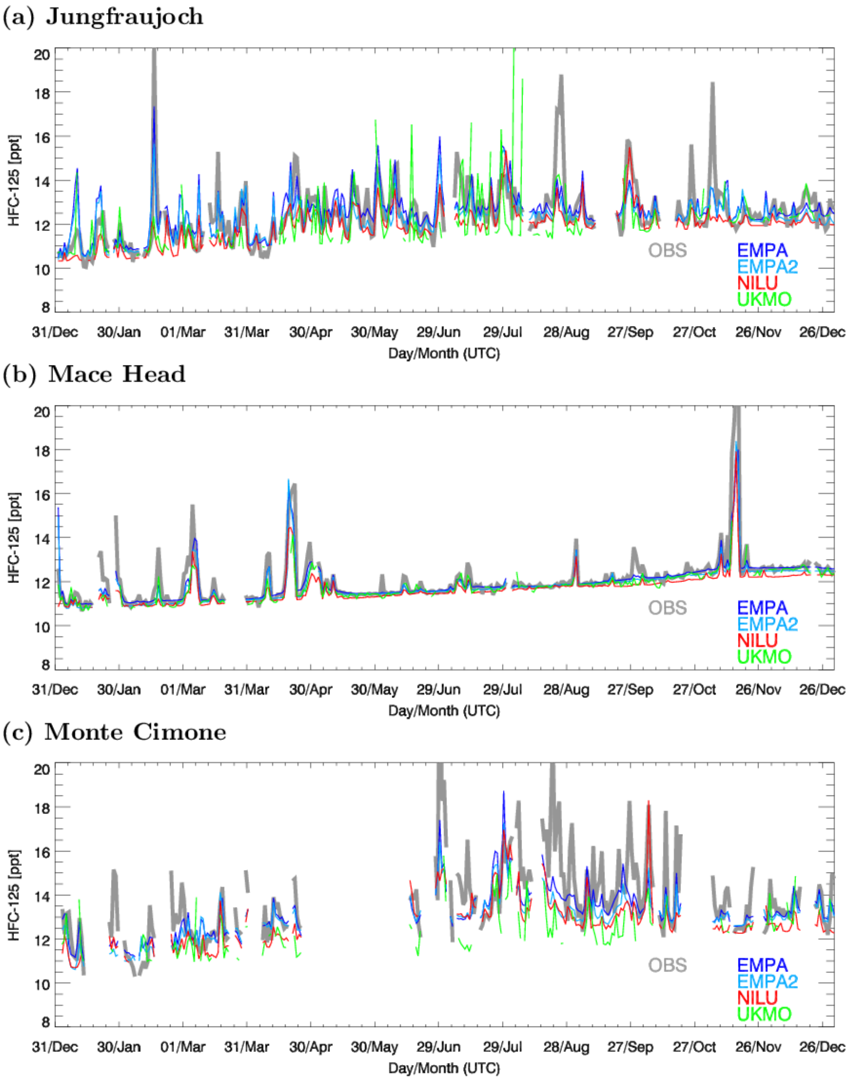
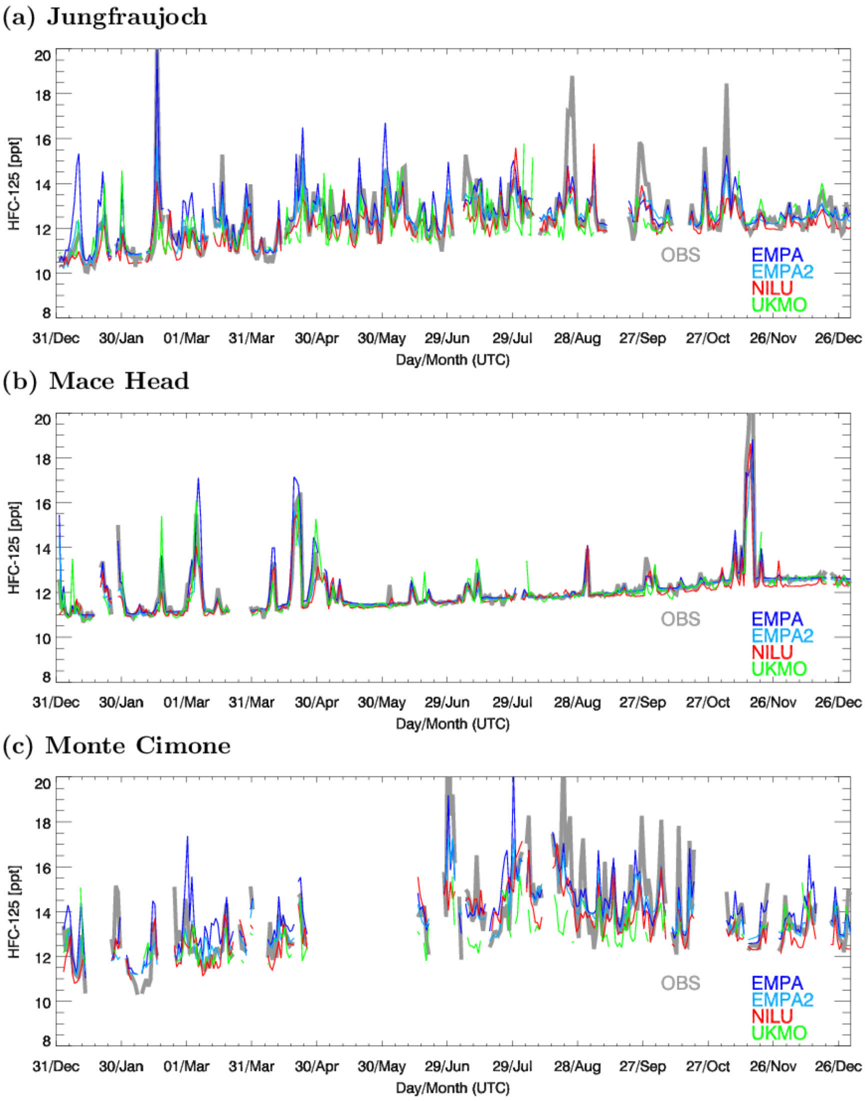


Figure 2: Prior simulated HFC-125 mole fractions (colour lines) overlaid over observations (thick grey line) at the three sites Jungfraujoch, Mace Head and Monte Cimone.

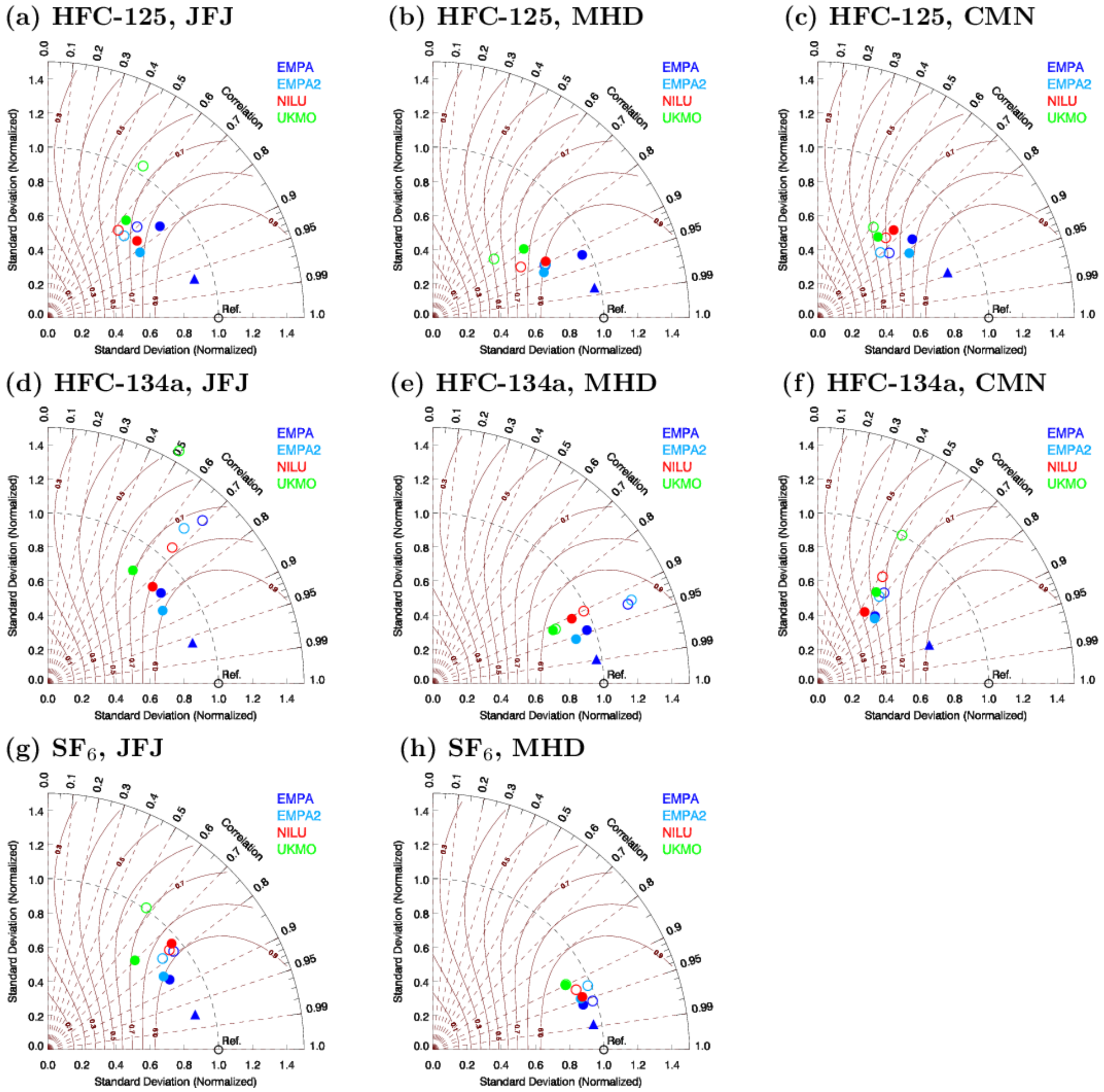


(a) Jungfraujoch

(b) Mace Head

(c) Monte Cimone

Figure 3: Same as Figure 2 but for posterior simulations.


(a) HFC-125, JFJ

(b) HFC-125, MHD

(c) HFC-125, CMN

(d) HFC-134a, JFJ

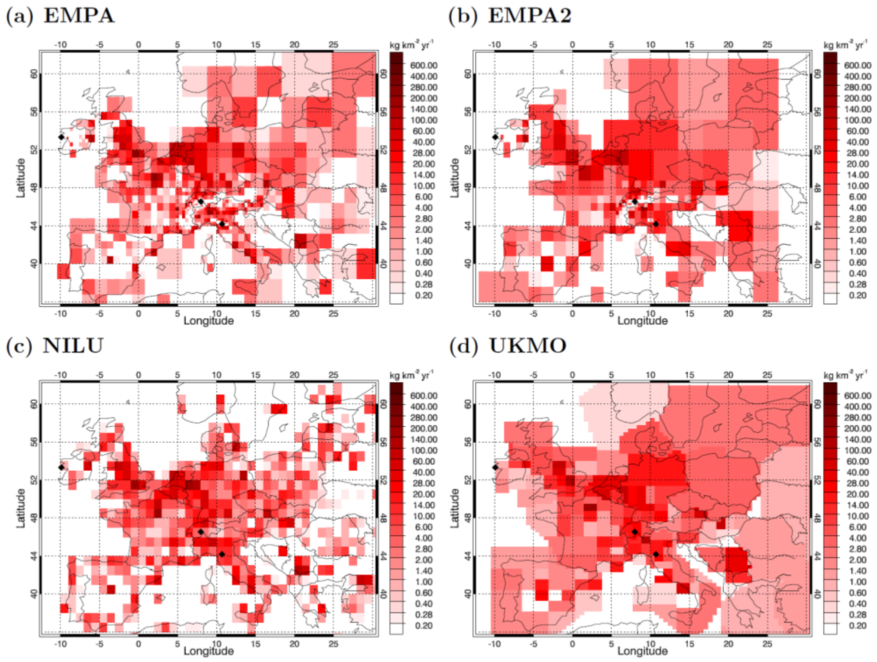
(e) HFC-134a, MHD

(f) HFC-134a, CMN

(g) SF_6, JFJ

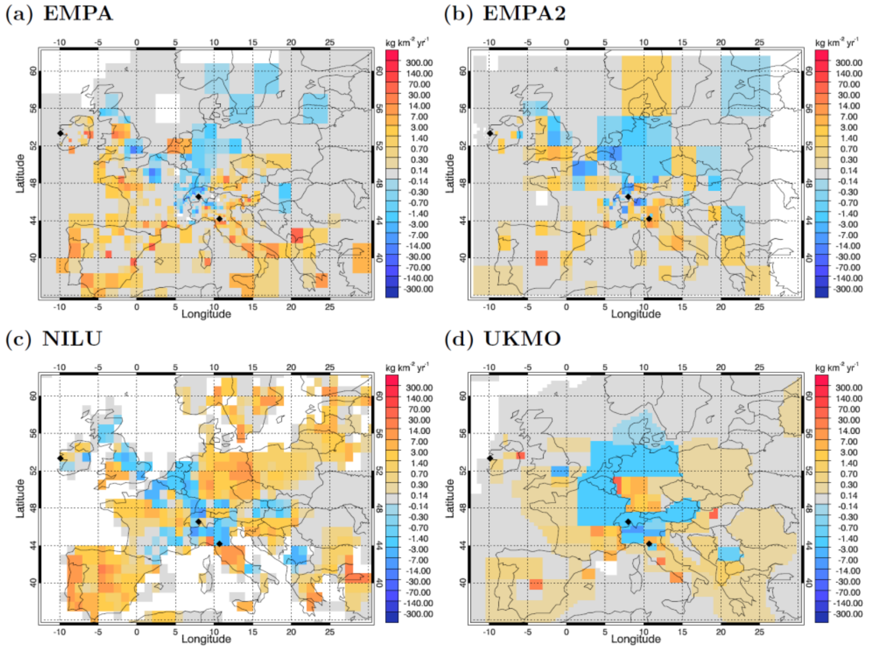
(h) SF_6, MHD

Figure 4: Taylor diagrams of model performance for the simulated prior (open circles) and posterior (filled circles) mole fraction time series. The filled blue triangle for EMPA indicates the performance when including an AR(1) autocorrelation term in the Kalman filter. The linear distance from the reference point (Ref.) is proportional to the centred (bias corrected) root mean square error (RMSE). The angle of rotation with respect to the vertical axis corresponds to the Pearson correlation coefficient R.



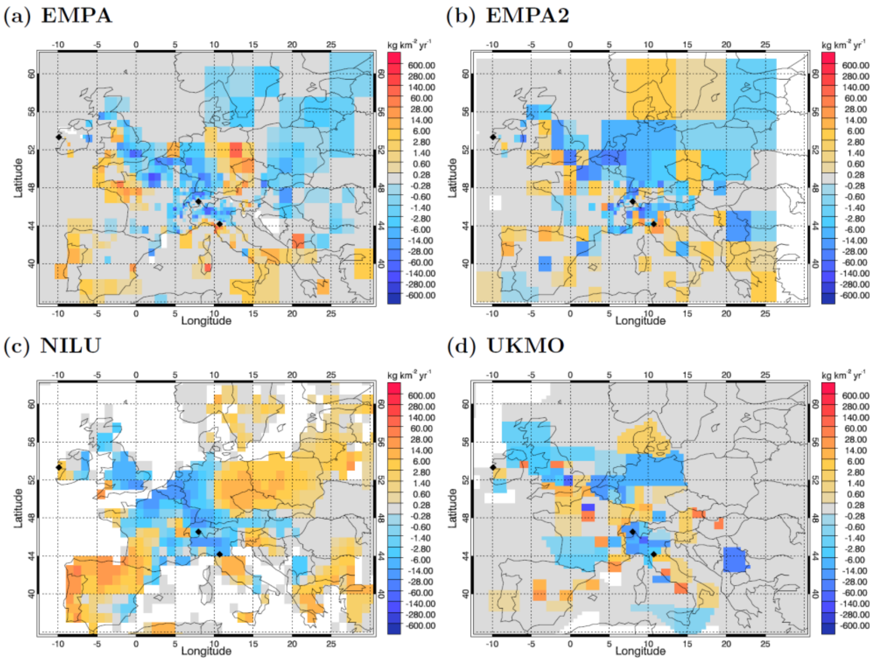
(a) EMPA (b) EMPA2 (c) NILU (d) UKMO

Figure 5: Prior emissions of HFC-134a as represented in the four inversion systems.



(a) EMPA (b) EMPA2 (c) NILU (d) UKMO

Figure 6: Posterior – prior HFC-125 emission differences (experiment M1).

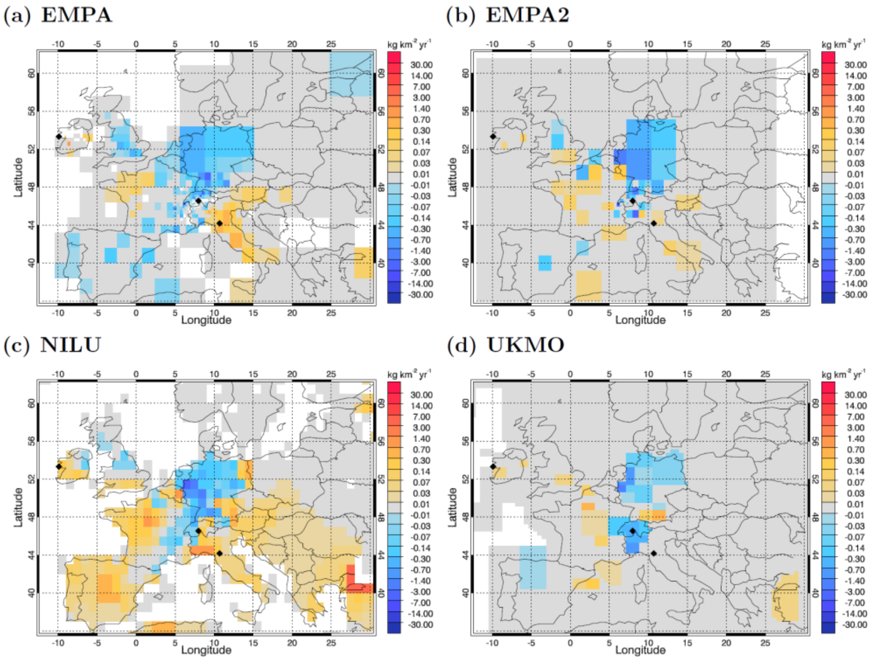

(a) EMPA

(b) EMPA2

(c) NILU

(d) UKMO

Figure 7: Posterior – prior HFC-134a emission differences (experiment M2).

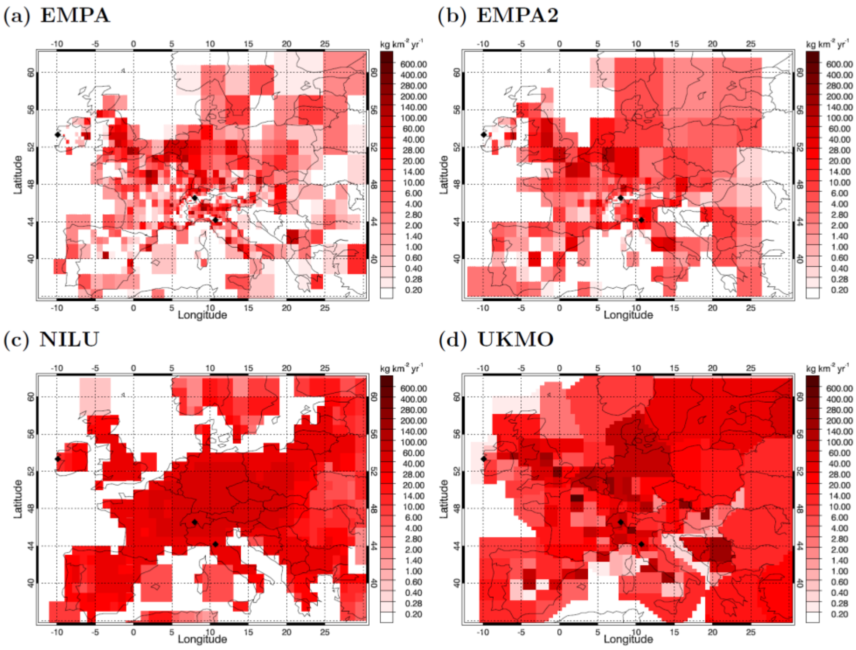

(a) EMPA

(b) EMPA2

(c) NILU

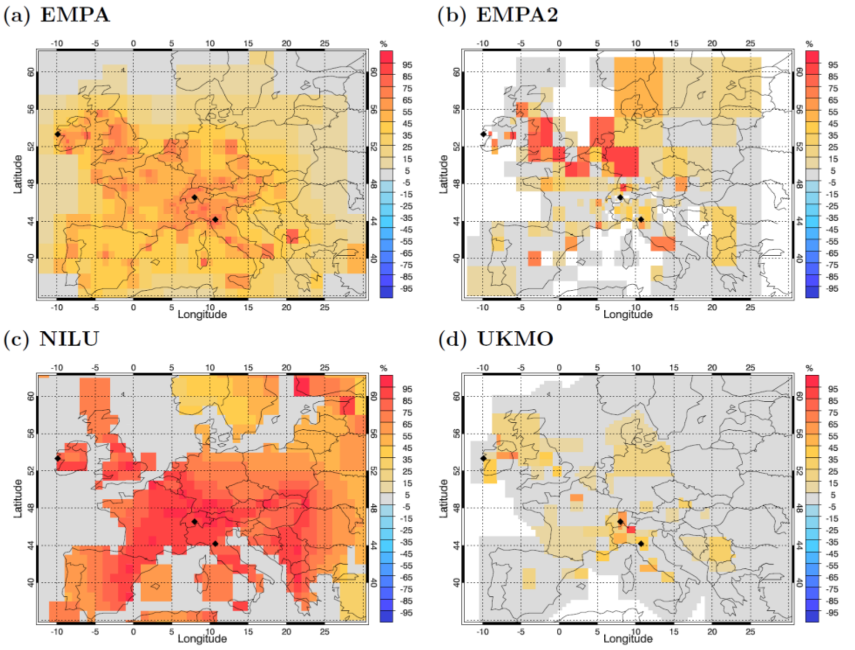
(d) UKMO

Figure 8: Posterior – prior SF_6 emission differences (experiment M3).



(a) EMPA (b) EMPA2 (c) NILU (d) UKMO

Figure 9: Uncertainty of prior HFC-134a emissions (experiment M2).



(a) EMPA (b) EMPA2 (c) NILU (d) UKMO

Figure 10: Uncertainty reduction (1-u_{post}/u_{prior}) in % for HFC-134a (experiment M2). For EMPA, the reduction is shown in terms of reduction of relative uncertainties [1-(u_{post}/x_{post})/(u_{prior}/x_{prior})].

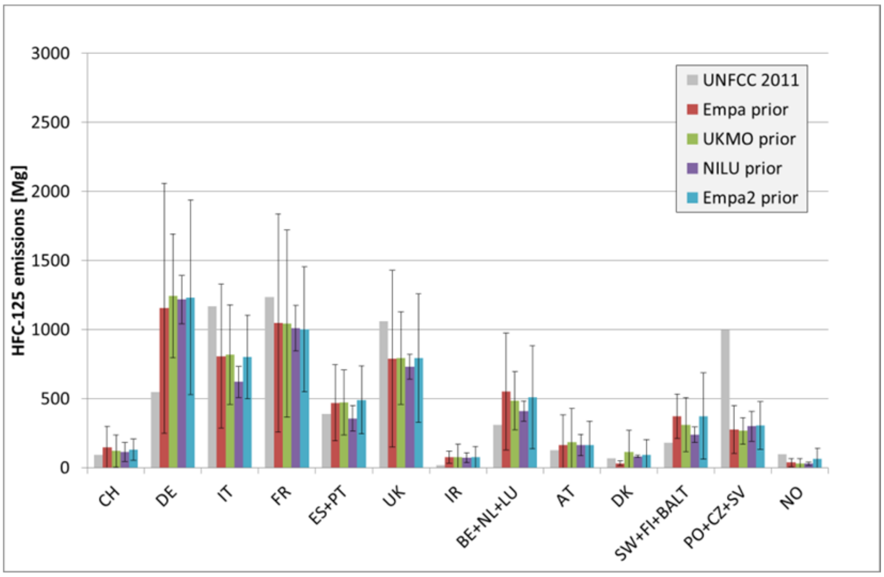


Figure 11: Country-aggregated prior emissions of HFC-125 (experiment M1). Country codes following ISO2 conventions except for BALT = Baltic countries (Estonia, Latvia and Lithuania). CH=Switzerland, DE=Germany, IT=Italy, FR=France, ES=Spain, PT=Portugal, UK=United Kingdom, IR=Ireland, BE=Belgium, NL=Netherlands, LU=Luxemburg, AT=Austria, DK=Denmark, SW=Sweden, FI=Finland, PO=Poland, CZ=Czech Republic, SV=Slovakia, NO=Norway.

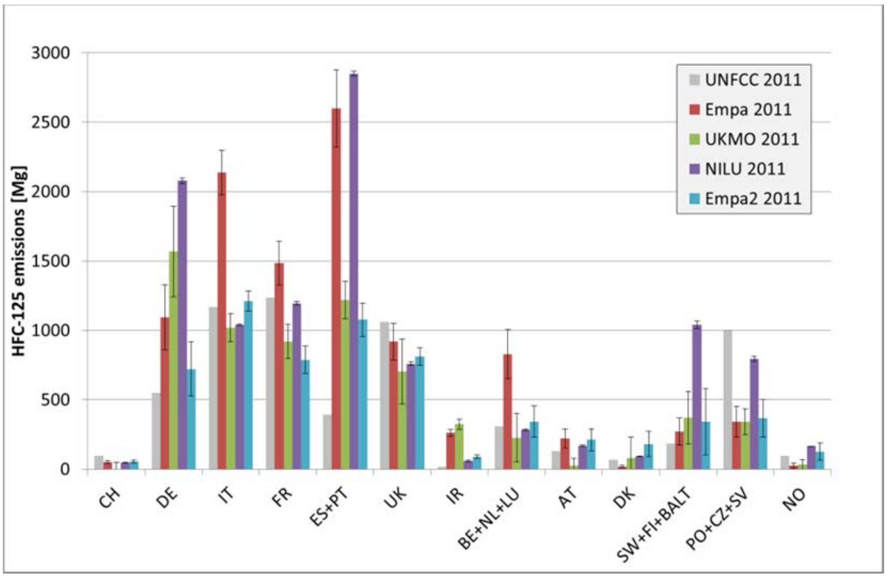
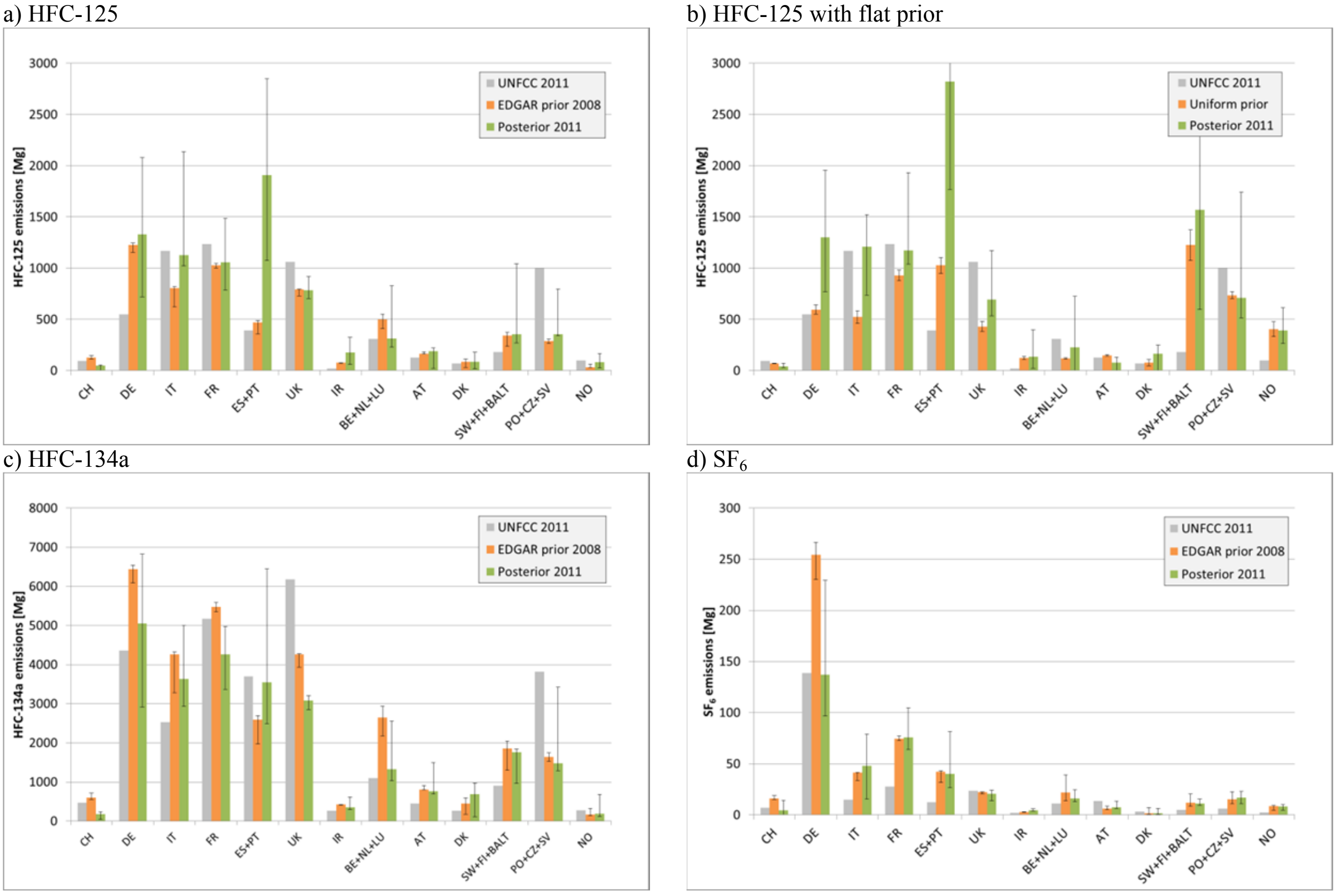


Figure 12: Country-aggregated posterior emissions of HFC-125 (experiment M1).

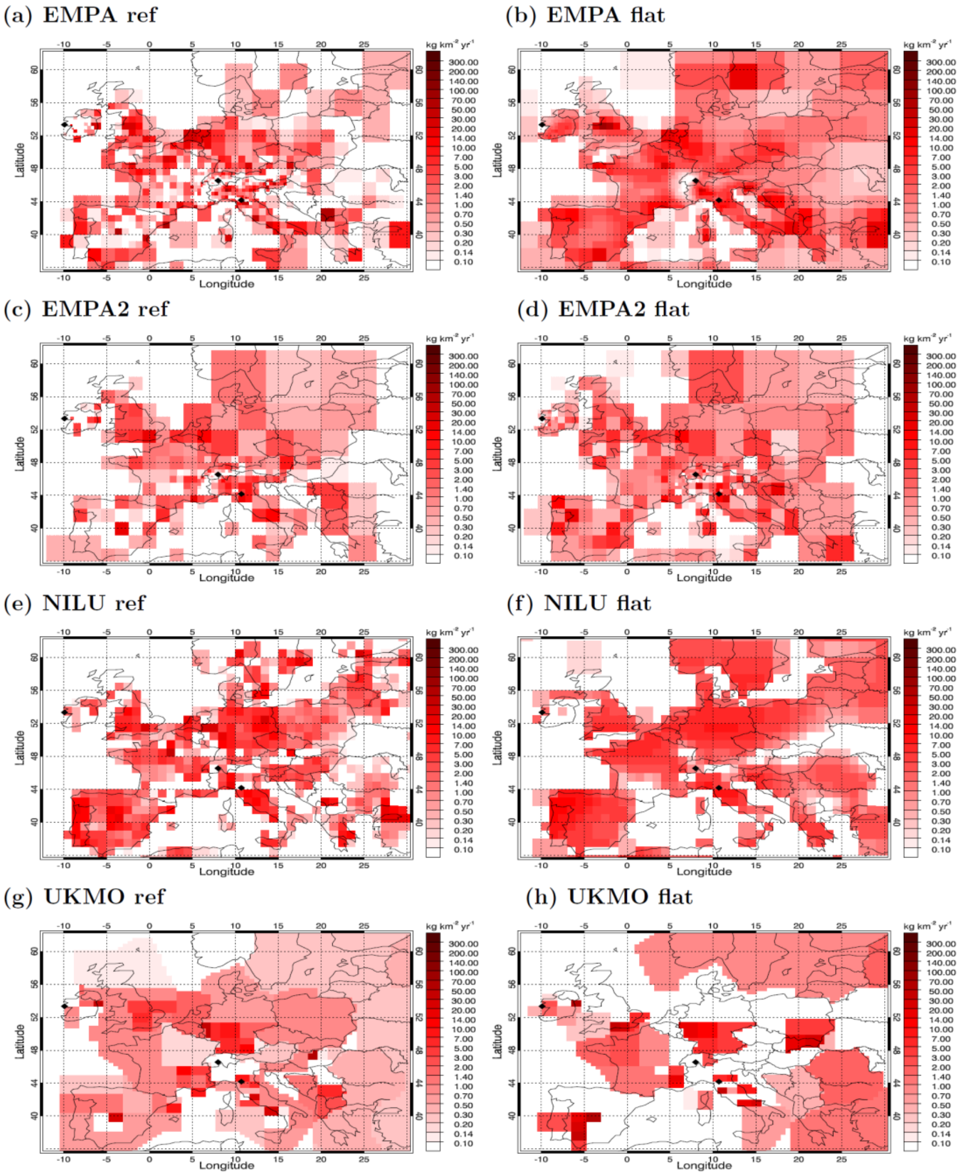

a) HFC-125

b) HFC-125 with flat prior

c) HFC-134a

d) SF_6

Figure 13: Median country-aggregated posterior emissions for a) HFC-125 (experiment M1), b) HFC-125 with flat prior (experiment FLAT), c) HFC-134a (experiment M2), d) SF_6 (experiment M3). Uncertainties bars denote the range between minimum and maximum estimate of the four models.


(a) EMPA ref

(b) EMPA flat

(c) EMPA2 ref

(d) EMPA2 flat

(e) NILU ref

(f) NILU flat

(g) UKMO ref

(h) UKMO flat

Figure 14: Posterior emissions of HFC-125 for the reference experiment M1 (left column) and the experiment with flat prior (right column).