Author's response #1

This paper presents a study into two factors that influence inverse estimates of greenhouse gas emissions using Lagrangian particle dispersion models (LPDMs): the period over which the models are run backwards in time, and the choice of "baseline" estimation method. I think this paper is suitable for publication in GMD. However, I think greater care needs to be taken when attempting to generalise some results.

My main criticism of the paper is around the way that the use of backward simulation time is discussed. It is no doubt true that very short simulation lengths will likely "miss" important influence on observations of nearby sources. However, there will be substantial diminishing returns for very long simulation periods, because of the rapid decline in sensitivity and increase in model uncertainty with distance from the measurement stations (and hence simulation time). Therefore, I think the authors need to be careful not to imply that we can solve for very far-field emissions using a very sparse measurement network and "long-enough" simulation periods. To provide some context, a recent study (Rigby et al., 2019) that used Gosan and Hateruma data suggested that robust emissions estimates of CFC-11 could only be derived for the eastern provinces of China (10 and 30-day simulations were used), and that as additional provinces were added further west, a posteriori uncertainty increased dramatically (and I believe the level of agreement between the four different methods decreased). In this paper, it is implied that, using essentially the same network in Asia, we could infer SF6 emissions from, for example, India (thus extending the domain from ~100s of km to ~1000s of km) by going from 5 to 50-day simulations. While there is of course some sensitivity to these distant regions, in practice, I very strongly doubt whether robust emission estimates could be derived so far from the stations, because of the very small SRR (around two orders of magnitude difference between India and eastern China, according to Figure 2) and comparatively high model uncertainty for such long trajectories.

Building on this point, I believe that one of the reasons this framing is arrived at is because the GDB method gives consistent results for different simulation lengths (Figure 10). However, I assume that the reason for this invariance to simulation length is because the inversion does not significantly deviate from the prior in regions further from the measurement station (at least when integrated over large scales). In the case of very short simulations, the sensitivity is necessarily only to emissions very close to the measurement stations. But even in the case of very long simulations, the SRR is so small in far regions, compared to the near-by areas, that it doesn't really make much difference at the global scale. Indeed, I think that this explains the persistent bias for all simulation lengths when the a priori fluxes are changed in Figure 14 (i.e., the low sensitivity prevents the inversion from overcoming the bias in the prior). I do think the tests show convincingly that the GDB method does a better job of compensating for issues relating to simulation length than the other baseline methods. Therefore, I suggest the authors take care that the discussion emphasises this outcome, without implying that more can be drawn from very long simulations than is possible in practice. Further elaboration and specific suggestions are provided below.

We would like to thank reviewer #1 for the valuable and constructive review of our manuscript.

In the response we use 4 different colors. The blue colored text is the general answer to the reviewer's comments. Additionally, we show how the text is changed in the manuscript: The original text is colored grey, removed text is colored red, and new text is colored green.

We acknowledge that the discussion around the backward simulation period needs to be handled with more care. We totally agree with the fact that a lack of observations in a sparse measurement network cannot be compensated through “long enough” simulation periods and it is not our intention to imply that. It is correct that the SRR values get more widespread and thus less specific with every additional day of backward calculation and therefore the emission patterns in certain regions far away from the observations network cannot be determined accurately, even with relatively long simulation periods. At the same time, it is clear that one should make the best use of the available observation data, and we suggest that there is still valuable information contained in the SRR values for backward simulation times longer than 5-10 days, which increases the value of the available data. We support this claim with three facts:

1) The correlations between modeled and observed mixing ratios (Table 2; averaged over all stations) continuously increase with increasing backward time. The increases in the explained variance per day backward becomes smaller and smaller with time, so is incremental but can be noticed even up to 50 days. Thus, to make the best use of the existing observations, longer backward simulations are beneficial, even though the gain in information decreases with every day backward. As long as the costs for longer backward calculations are small compared to the costs involved in performing the measurements, the diminishing return in improved accuracy for every extra day of calculation should not discourage us from making longer simulations.
2) Longer backward simulations improve global emission estimates as shown in Fig. 14, even though the improvement of regional emission patterns will be limited. Regarding Fig. 14, we think that an extension of the backward simulation period beyond 50 days would further reduce the bias of the retrieved global emissions. Imagine the extreme case of multi-year-long simulations: particles will be equally distributed around the globe and the total global emissions could be derived similar to a box model - even with only one station. Thus, an inversion based on very long backward simulation periods should always give a quite accurate global emission, presumably with a better accuracy than a simple global box model - at least for species with lifetimes of decades or more. Inversions with shorter backward simulation times, in contrast, will usually result in erroneous global emissions. Fig. 14 confirms this behavior for up to 50 days but we have no doubt that even longer simulations would give even more accurate global emissions than we have obtained with our 50-day simulations, especially when starting with a “wrong” prior.
3) We argue further that by getting a better constrain on the global emissions, and on regions which are well covered by observations, we also get improved emission estimates of poorly sampled regions, however, without resolving the exact spatial emission patterns in these poorly sampled regions. We think that emissions can be improved as long as the growth in SRR is above linear (Fig. 11), which for some regions will happen only after a very long simulation period. As shown in Fig. 12, the inversion produces non-zero emission increments at regions, that were untouched for smaller simulation periods. We would like to repeat that we do not claim that the regional emission patterns in poorly sampled continents like South America will be very accurate – there simply is not enough information available there, as the reviewer correctly points out. However, by having accurate emission information in well sampled regions and an additional strong constraint on the global emissions (which inversions with shorter backward simulations do not have), emissions in poorly sampled regions will also be corrected. Like global box model variants (e.g. AGAGE 8-box model) are able to retrieve emissions in several latitude bands, long FLEXPART simulations should also be able to give some information on the location of the emissions, even though the resolution will be very coarse.

Therefore, there should be enough information in the simulations that the inversion does not distribute the residual emissions (global minus well-constrained ones) randomly on the globe but still has enough skill to attribute them to the correct hemisphere, and hopefully also to the right continent at least. We discuss this in the text (see comment to L12, Final remark)

Similarly, care needs to be taken in the discussion of how different choices of baseline method should be used. There are examples where statistical baselines (perhaps including some baseline optimization) have provided consistent results to methods similar to GDB (e.g., Rigby et al., 2019; Brunner et al., 2017). Therefore, I think it is too broad to draw the conclusion that these methods need to be "abandoned" (Line 517). Rather, perhaps the case needs to be made that careful consideration should be given to the type of problem in which they are used.

We agree that our conclusion was perhaps too strong and have changed that in the text (see reply to the comment to L517).

Specific comments:

L6: suggest rewording: "… that purely statistical baseline methods CAN cause large systematic errors" (note "systematic", rather than "systematical")

- changed accordingly

L7: suggest removing "highly" before sensitive

- changed accordingly:

L8: In the final part of this sentence, I don't think it's quite fair to say "and that are consistent with recognized global total emissions". As discussed, I feel that agreement is primarily because the "recognised" emissions are used as the prior, and therefore these prior values are retrieved for integrated areas far from the measurement sites either because the particles haven't reached them yet (short simulations) or because the SRR is very low (even for long simulations). Of course, by design, and as demonstrated, the GDB method does a better job at accounting for the "missing" SRR in short simulations.

- We agree, that the good agreement for all simulation periods is partly due to the prior, especially for short backward simulation periods. In Fig. 14 we show the case of a strongly biased prior, where inversion results deviate strongly from the known global totals for short simulation periods but improve with longer periods. We think it is still important to mention that with the GDB method, global emissions stay close to the global prior, which is not the case for the other investigated methods.

and that are consistent with recognized global total emissions → and that show a better agreement with recognized global total emissions

L12: "Further, longer periods help to better constrain emissions in regions poorly covered by the global SF6 monitoring network (e.g., Africa, South America)". As discussed, I don't think this case has been made in this paper. I think this sentence should be removed.

- As mentioned before, we think the statement is actually true, but we weakened it and address your concerns in more detail in the discussion part.

Further, longer periods help to better constrain emissions in regions poorly covered by the global SF_6 monitoring network (e.g., Africa, South America) → Further, longer periods might help to better constrain emissions in regions poorly covered by the global SF_6 monitoring network.

- We added L488:
  - Final remark

In this study, we show many advantages for using relatively long backward simulation periods for the inversion. Nevertheless, the improvement of regional emission patterns is still limited by the observation network. A lack of observations in one region cannot simply be compensated by extending the simulations for stations in other regions to very long periods. For backward simulation times of 20-50 days, the emission sensitivity is distributed over large areas but usually still concentrated within broad latitude bands. The additional information to be gained from such long simulation times, on top of the information provided by the shorter simulation times, can probably best be compared with the inversions done with a multi-box model such as the AGAGE 8-box model (e.g. Rigby et al. ;2013) that is capable of determining the emissions in broad latitude bands. Consequently, if the emissions in certain regions with a dense observation network are already well constrained by shorter simulation periods, the residual emission will be attributed correctly as an emission total to all other regions of the same latitude band with a poor station coverage. The effective resolution of the obtained emissions in such data-poor regions may be very coarse but the result might still be informative. Furthermore, the emission sensitivity for the 20-50 day backward period is still not uniformly distributed over a latitude band and thus provides some limited regional information. Perhaps supported with a limited number of strategically located flask measurements, inversions with long backward simulation times could provide coarse but robust information on emissions in poorly sampled regions. Independently, the growing correlation between modeled and observed mixing ratios with increasing backward simulation length (Table 2; averaged over all stations) also shows that longer backward simulations hold additional information, even though the information gain decreases with every day added to the simulation length and probably becomes marginal for very long backward simulation times. However, we propose to make use of this additional information and apply longer periods whenever possible to make the best use of the existing observation network.

L33: “frequency”, rather than “frequent”

- changed accordingly:

L37 – 39: Care needs to be taken with the periods ascribed to each study here. E.g., Brunner et al. (2017) uses LPDM runs from 5 to 19 days in length (see, Table 1 in that paper), not just 5 days; Rigby et al. (2019) has LPDM runs of 10 or 30 days, not just 10 days.

- changed accordingly:

…. that they are often run backward in time only for a few days ….

Brunner et al., 2017 → Vollmer et al., 2009

Rigby et al., 2019 → Thompson et al., 2017

L60 – 68: Other approaches should be cited here. E.g., Hu et al. (2019) compared a GDB-type approach to statistical methods, Lunt et al. (2016) uses a GDB method that uses mole fraction “curtains” around a regional domain (e.g., the termination points are tracked in space, rather than time), Rigby et al. (2011) and Ganshin et al. (2012) developed a nested Eulerian/Lagrangian approaches. It should be emphasised that in many of these papers, the “baselines” are adjusted as part of the inversion (so consider adding to the list on L155).

- changed accordingly:

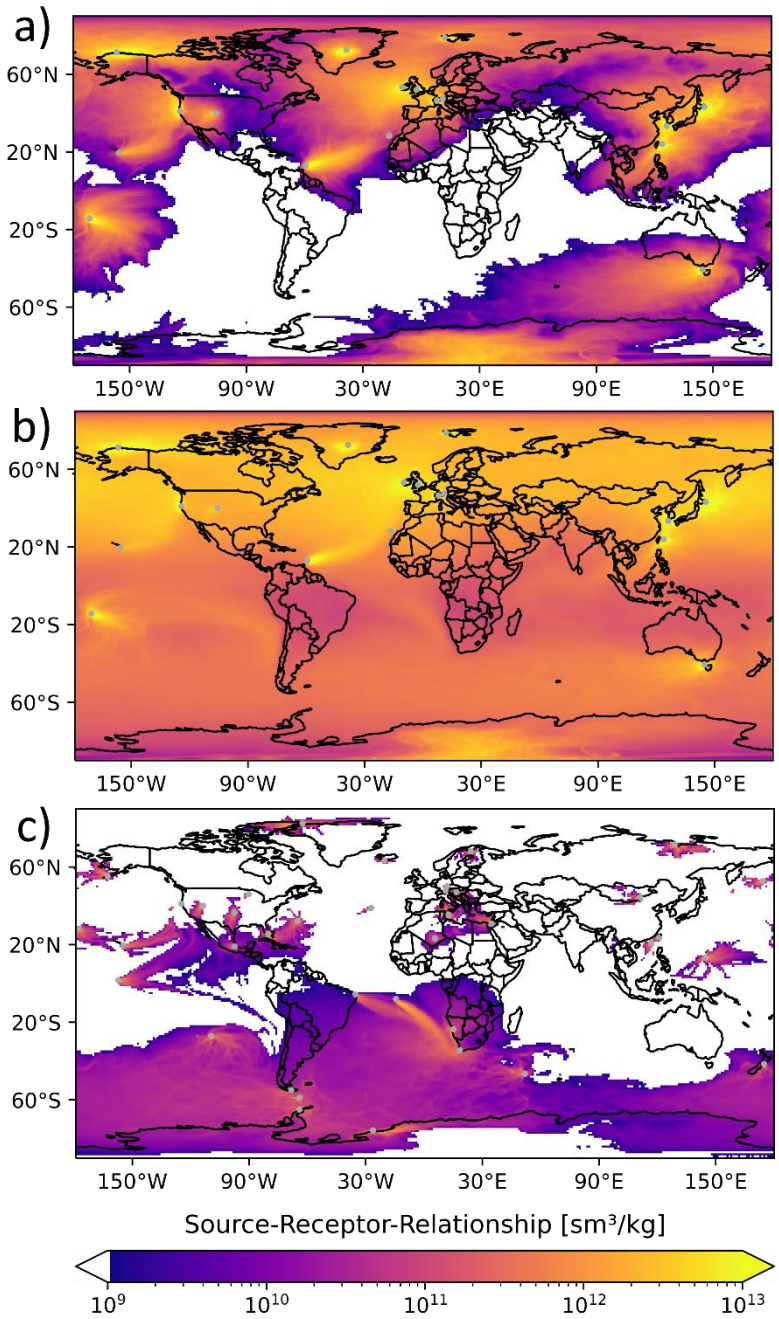
Apart from using observations at each individual station to maintain a baseline, Rödenbeck et al. (2009) suggested a general "nesting" scheme, where a regional transport model – either a Eulerian or Lagrangian model – is embedded into a global model providing information from outside the spatio-temporal inversion domain. Such a global distribution based (GDB) approach was used by e.g. Trusilova et al. (2010) and Monteil and Scholze (2021) for carbon dioxide, and similar by Thompson and Stohl (2014) for methane. Whereas Rödenbeck et al. (2009) coupled the LPDM back-trajectories with the global model in the space domain, Thompson and Stohl (2014) did the coupling at the time boundary. →

Apart from using observations at each individual station to maintain a baseline, Rödenbeck et al. (2009) suggested a general "nesting" scheme, where a regional transport model – either a Eulerian or Lagrangian model – is embedded into a global model providing information from outside the spatio-temporal inversion domain. Such a global distribution based (GDB) approach was used by many authors: Trusilova et al. (2010) and Monteil and Scholze (2021) used Rödenbeck’s approach to estimate CO_2 emissions. Similarly, Rigby et al. (2011) and Ganshin et al. (2012) developed approaches to nest a Lagrangian into a Eulerian model and tested it for SF_6 and CO_2, respectively. Estimating CO_2 baseline mole fractions for inverse modeling, Hu et al. (2019) applied two GDB approaches and a statistical method, where a subset of observations with minimal sensitivity was selected to correct a GDB baseline. Lunt et al. (2016) and Thompson and Stohl (2014) applied GDB approaches to model CH_4 While Thompson and Stohl (2014) coupled the LPDM back-trajectories with the global model at the end of the trajectories (which are terminated after a defined time), Lunt et al. (2016) used the exit location of the particles leaving the inversion domain for the coupling.

Added to list at L155: Rigby et al. (2011)

Figure 2: How have SRR due to flasks and high-frequency sites been combined here? Since they have a very different frequency, has there been any effort to “weight” their influence? If not, it may be worth adding this as a caveat. I.e., even though the flask footprint might look quite substantial, it corresponds to very few data points (and therefore relative influence on an inversion).

Indeed, for the figure the samples were not weighted by the number of observations at individual sites. However, we changed the figure in the revised manuscript to account for the variable sampling frequency at different sites, by weighting the sensitivities with the respective observation number.



Caption Figure 2: c) shows the SRR for the case of using surface flask measurement sites in addition to in situ measurements and for a 50 day simulation period. → c) shows the increase in the annual averaged SRR due to the use of flask measurements in addition to continuous measurements for the case of a 50-day backward simulation period.

L181: When also using surface flask measurements (Fig. 2c) in addition to in situ measurements for the case of a 50 day backward simulation period, the emission sensitivity is substantially higher almost everywhere and more smoothly distributed over the globe. However, regions of low sensitivity remain in the Tropics and in the Southern Hemisphere. → Figure 2c shows the increase in the annual averaged SRR due to the use of flask measurements in addition to continuous measurements in the case of 50-day simulations. One can see substantial increases in the vicinity of the measurement sites, that quickly decline with distance to the sites. Further SRR values increase in large parts of the Southern Hemisphere, however, the increases over southern continental areas are relatively low, as most flask measurements are not well located for inversion purposes.

L241: The use of “eliminate errors” and “any bias” is too strong. You can’t eliminate errors or bias.

- changed accordingly

… on the ability to eliminate errors, and especially any bias of the global 3D mixing ratio fields … → … on the ability to minimize errors, and especially bias of the global 3D mixing ratio fields …

L245: Second sentence of this paragraph needs rewording, as it’s not clear what the “It” refers to.

- changed accordingly

… provide a detailed description of FLEXPART CTM and evaluate it for the example of methane.
… provide a detailed description of FLEXPART CTM and evaluate this model for the example of CH4.

L263: Suggest “A priori emissions”, rather than “information”

- changed accordingly:

L267 – 268: as noted in the preamble, it should be noted here that this choice of prior introduces some “circularity” into some of the results.

- we added:

Note at this point that the a priori emissions as constructed agree with recognized global emissions, which should be kept in mind when the global total is used as a reference value in the discussion.

Section 3.1: I think it needs to be mentioned here that both of these stations are somewhat complex in terms of baseline estimation, in that they periodically intercept air from the southern hemisphere. As noted, at Gosan, the summer months are characterised by southern hemisphere baselines. Therefore, I think these stations are likely to be among the most challenging in the world for statistical baseline estimation. I think the investigation would be improved if a station with less complex baseline were also added. For example, are similar results obtained if Mace Head is used?

- We chose these stations, as they are ideal to discuss the differences between the three investigated baseline methods. We agree however that Gosan and Barbados are challenging and that statistical methods might work better at less complex ones. We will discuss this in the text and refer to figures in the supplement, where we show the timeseries for all stations. We added:

L281: Both, Gosan and Ragged Point periodically intercept air from the southern hemisphere and therefore have a rather complex baseline.

L363: On the other hand, statistical baseline methods might work better at observation stations, where the baseline termination is less complex. At Mace Head (Fig. S18) for

instance, both REBS and Stohl's method lead to a very high correlation between modeled and observed mixing ratios for the case of a 50-day backward simulation (r^2=0.87). Nevertheless, for the REBS method, the discussed growing negative bias with longer simulation periods can be observed.

L287: "… as a result the baseline should become lower and smoother in order to leave a priori mixing ratios unchanged." I don't think this statement applies to REBS (it's presented as applying to all methods)

- The statement was meant in the sense, that this would be desirable (for a good baseline method)

Ideally, the choice of the backward simulation period should have no systematic effect on the calculated a priori mixing ratios. By increasing the backward simulation time, and therefore enlarging the temporal domain, more direct emission contributions are included. All these direct emission contributions should be removed from the baseline and as a result the baseline should become lower and smoother in order to leave a priori mixing ratios unchanged. Furthermore, one can assume that a correctly working baseline method leads to a proper agreement between *a priori* mixing ratios and observations. This agreement is investigated here for the three methods with → Ideally, the choice of the backward simulation period should have no systematic effect on the calculated a priori mixing ratios. By increasing the backward simulation time, and therefore enlarging the temporal domain, additional emission contributions are included in the optimization. Per definition, these contributions are not part of the baseline and should ideally be removed from it. As a result, the baseline should become lower and smoother when the simulation period is increased. We investigate the agreement between modeled and observed mixing ratios for the three methods with ….

L289: "… leads to a proper agreement between a priori mixing ratios and observations". I'm not sure what this means (the use of the term "proper").

- changed accordingly, see L287

L295: suggest removing "… when direct emission contributions get more impact", as it's not clear what this means, and seems to be unnecessary.

- changed accordingly

L307: I think it's too strong to say that Ragged Point is "uninfluenced" by polluted air masses. Pollution events are observed. They just tend to be very small (and/or well captured by short LPDM runs, since any sources are likely to be very local). Also note that L314 seems to conflict with this line, because it references an increasing direct emission contribution.

- changed accordingly:

Since Ragged Point is uninfluenced by regional emissions, no significant measurement peaks need to be excluded → Since regional pollution events captured at Ragged Point tend to be very small, no significant measurement peaks need to be excluded

L332: "… can only reproduce a few pollution events at Gosan…". Is this demonstrated in the 0-day simulation? If so, say so explicitly.

➢ Yes, we changed accordingly:

… it can only reproduce a few pollution events at Gosan, underestimates the highest and overestimates the lowest measured SF_6 mixing ratios (Fig. 6a). → … it can only reproduce a few pollution events at Gosan, underestimates the highest and overestimates the lowest measured SF_6 mixing ratios, as demonstrated in the 0-day case (Fig. 6a).

L334: "…provides, in principle, infinite resolution". I suggest this should be reworded, as infinite resolution isn't possible (for computational reasons and the resolution of the meteorology).

➢ changed accordingly:

…provides, in principle, infinite resolution → … provides much higher resolution

L340: Suggest rewording to: "… reproduces the measured mixing ratios well. However, it generates more variability than observed at this station"

➢ changed accordingly

L352: "Neither the REBS nor Stohl's method could correctly reproduce these negative SF6 excursions". As noted on the overarching comment for this section, this isn't surprising, as these methods aren't really designed for such complex baselines.

➢ Yes, it is not surprising that the statistical baselines do not work well for these stations. However, studies have applied these statistical methods to Gosan and Barbados (e.g. Fang et al. 2012, Stohl et al. 2009, 2010, Vollmer et al., 2017).

L354: "… it reproduces measurements insufficiently…". Not sure what this means. Do you mean the simulation can be biased?

➢ changed accordingly:

Despite of all advantages of the GDB method, it reproduces measurements insufficiently if the modeled global mixing ratio fields are biased. → Despite of all advantages of the GDB method, it doesn't work well if the modeled global mixing ratio fields are biased.

L367: Remove comma after "surprising"

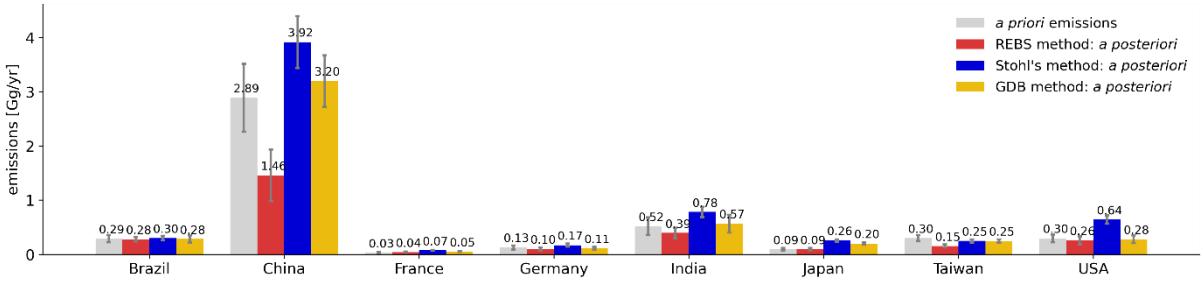
➢ changed accordingly

L371: Remove "This is quite a substantial improvement." This is a subjective judgement. Leave this up to the reader to interpret.

➢ changed accordingly

Figure 9: Show posterior uncertainties.

➢ changed accordingly:



Caption Fig.9: National SF_6 emissions for selected countries, based on 20 day LPDM backward calculations with different choices of the baseline method. Uncertainties represent a 1σ range.

L384 – 392: There's an implicit assumption in this paragraph that the GBD method is "just right", and that the other methods are too high or too low. At this stage, we can't be sure which one is right or which is wrong. We can only compare one method with another.

- Yes, we agree, as the real emissions are unknown, we cannot be sure, what is right or wrong. However, when considering the increments together with the biases in Tab.2, we think there is good reason to make this assumption. We therefore discuss your comment and make this clear in the text!

When using the REBS method (Fig. 8b), the inversion produces negative emission increments in almost all areas of the globe, indicating that calculated baselines are too high overall. This is consistent with the assumption that the method overestimates the baseline at individual stations by wrongly classifying observations as baseline observations that are actually influenced by emissions within the backward calculation period. In contrast, the inversion algorithm produces positive increments almost everywhere around the globe when applying Stohl's method (Fig. 8c), suggesting that the method systematically underestimates the baseline (not only at background stations) which generally leads to a priori emissions that are too high. In case of the GDB method (Fig. 8d) negative and positive increments are more balanced, showing no sign of a systematical under- or overestimation of the baseline. Large positive increments can be seen in East Asian regions and parts of Europe, whereas the inversion tends to produce slightly negative increments in the Southern Hemisphere. →

When using the REBS method (Fig. 8b), the inversion produces negative emission increments in almost all areas of the globe. As the real emissions are unknown, this is not necessarily an unrealistic result. However, when considering these mostly negative increments together with the discussed positive bias for REBS baselines in Table 2 (especially for longer backwards simulation periods), there is reason to assume that the REBS method overestimates baselines and consequently underestimates the *a posteriori* emissions overall. In contrast, the inversion algorithm produces positive increments almost everywhere around the globe when applying Stohl's method (Fig. 8c). Again, considering this together with the discussed negative biases in Tab. 2, this might indicate an underestimation of the baselines and an overestimation of the *a posteriori* emissions overall. In case of the GDB method (Fig. 8d) negative and positive increments are more balanced. Overall the patterns are more similar to the ones of the REBS method, except in East Asia, where they rather resemble the patterns of Stohl's method. Large positive increments can be seen in East Asian regions and parts of Europe, whereas the inversion tends to produce slightly negative increments in the Southern Hemisphere.

L396: Suggest re-wording: "…cases and therefore the baseline choice has little impact."

➢ changed accordingly

L399: “revealing systematic problems in the first two methods”. Again, now do you know that GDB is right and the others suffer from systematic problems?

➢ changed accordingly:

In almost all cases the REBS method leads to smaller and Stohl’s method to larger national emissions than the GDB method, again revealing systematic problems in the first two methods. Due to the large emissions in China these problems become especially apparent there →

In almost all cases the REBS method leads to smaller and Stohl’s method to larger national emissions than the GDB method. Due to the large emissions in China the differences in *a posteriori* emissions become especially apparent there ….

L404 – 405: I suggest noting again that this introduces some element of circularity (or at least note that this adds some nuance into the interpretation of the results)

➢ changed accordingly:

Notice that this is the same value used to calculate the a priori emissions, so the line represents also the global a priori emissions, which should be kept in mind for the interpretation of the results.

L424: This wording is too strong: “ability to ensure a flawless transition between the forward”, as it’s not possible to have a “flawless” simulation. But note also my suggestion in the preamble as to how one might interpret this result in terms of the low sensitivity of much of the world to the observations.

➢ changed accordingly:

Considering the inversion results based on the GDB method, global emissions are in good agreement with the box model result for all tested backward simulation periods. Furthermore, global emissions stay almost unchanged for different backward simulation periods, demonstrating again the method’s ability to ensure a flawless transition between the forward (Flexpart CTM) and backward calculation. →

Considering the inversion results based on the GDB method, global emissions are in good agreement with the box model result for all tested backward simulation periods, as the global *a posteriori* emissions stay close to the global a priori value. Furthermore, these global emissions stay almost unchanged for different backward simulation periods, demonstrating the method’s ability to adjust the baseline according to the sampled emissions of different simulation periods.

L433: I don’t think this is truly “exponential”

➢ changed accordingly (see comment to L433 – 434)

L433 – 434: “For these poorly-monitored countries only backward simulations beyond the usual 5-10 days used in most studies provide information for the inversion”. I disagree that 5 – 10 days is “usual”. But as I said at the beginning, I don’t think it’s correct to imply that we can gain valuable new information from this length of simulation.

- We weakened the statement accordingly:

For these poorly-monitored countries only backward simulations beyond the usual 5-10 days used in most studies provide information for the inversion. For these countries, the SRR increase with time flattens to a linear increase only for very long transport times, even beyond the 50 days used in this study. →

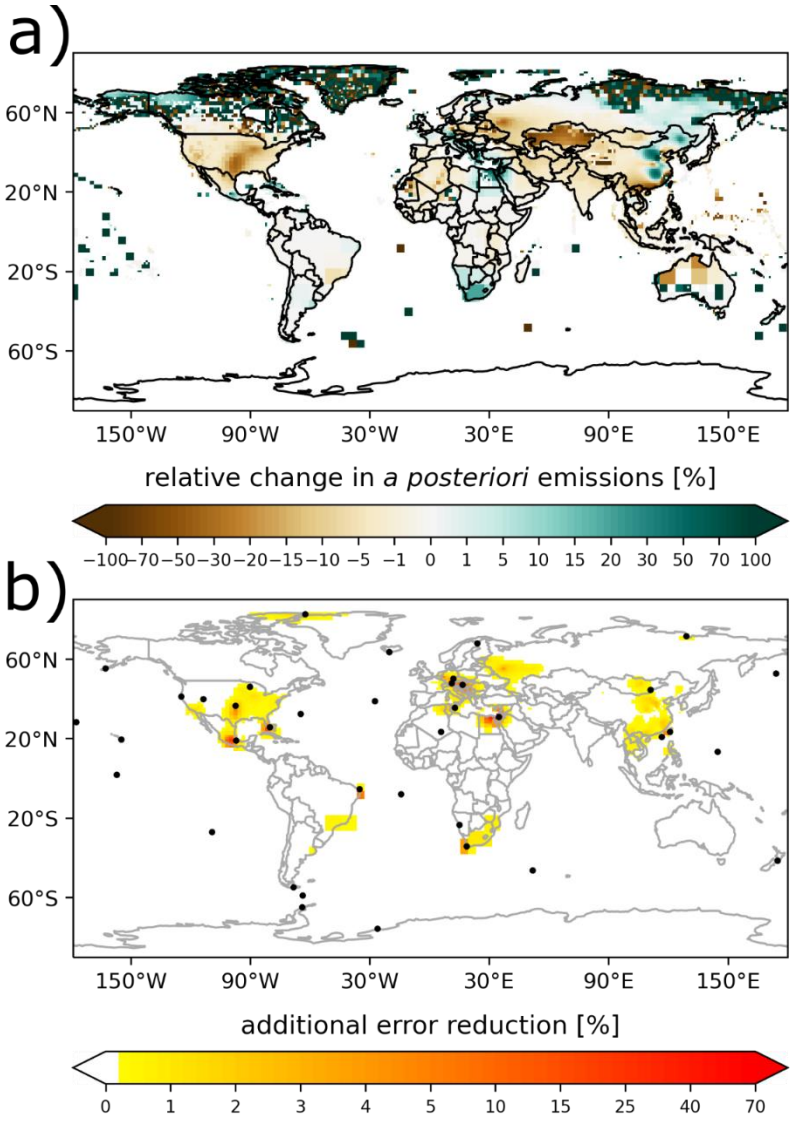
For countries poorly covered by the monitoring network, however, the SRR is close to zero for the first 5 to 15 backward days and only longer backward simulations might provide information for the inversion (see Fig. 11b). For these countries, the SRR increase with time flattens to a linear increase only for very long transport times, even beyond the 50 days used in this study.

Figure 13 caption: Some indication of the significance of this difference would be useful.

- We added a plot to Fig. 13 showing the additional error reduction due to the use of the flask measurements (we also changed the color bar to better distinguish it from Fig. 12,)

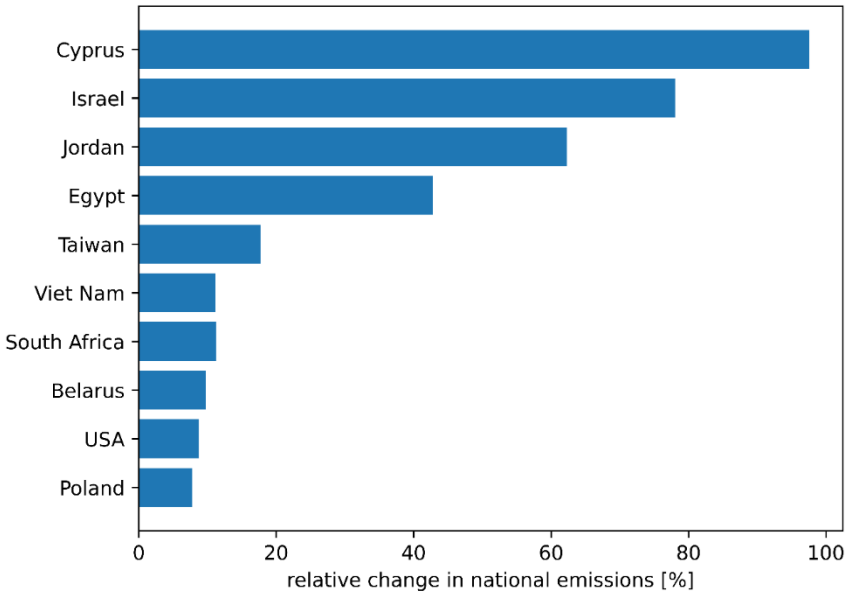
L450: “illustrating the great value of this additional information”. We need to see the uncertainties before we can decide if this demonstrates great value. Are any of these changes significant?

- Yes, we agree. We rewrote this section and discuss the changes together with the additional error reduction. Also, we added additional text and figure to discuss the impact of the additional flask data on the national and global emission estimates.



Caption Fig 13: a) Relative change in *a posteriori* emissions and b) the additional error reduction when using flask measurements in addition to in situ measurements for the 50-day simulation. The locations of the flask measurements are marked with black dots.

Figure 13a) shows the relative change in a posteriori emissions and Fig. 13b) the additional relative error reduction when using flask measurements additionally to the in situ measurements for the 50-day backward simulation. One can see substantial differences in the USA, Eastern Europe, South Africa, East Asia and the Near East, where also an additional error reduction occurs. While this additional error reduction can be relatively large (up to 73 %) for grid cells in the vicinity of the measurement sites, it quickly decreases down to a few percent with larger distance to the measurements. Consequently, flask measurements show only small influence on the total global emission estimate (< 1%), but can have a large impact on calculated national emissions of specific countries (Fig. A4). For countries in the Near East the additional use of flask measurements changes national emission estimates by 40 to 100%. South African and American emissions are modified by around 10%.



Caption Fig A4: relative change in national *a posteriori* emissions of selected countries, when flask measurements are used in addition to continuous measurements in the case of 50-day simulations.

L467: “However, it is clear that a substantial bias remains even with a backward simulation period of 50 days. It seems likely that an extension of the backward simulation period beyond 50 days would further reduce the bias.” As I said, I don’t think this is true. I think this likely comes from low sensitivity to emissions far from the measurement stations. Otherwise, the implication would be that we could overcome any bias in the prior using just one station and a very long simulation

- As mentioned above, we think this statement is actually true. In principle, even with one station and very long simulation periods, global emissions should be obtainable (similar to a box model). In our case, the total emission estimates should improve with increasing backward simulation time, however with varying spatial resolution of the emission patterns strongly depending on the observations network. We discuss this in the text (see comment to L12, Final Remark).

L494: remove “entirely”

- changed accordingly

L502: “… is superior and leads to a posteriori emissions that are far less sensitive to the LPDM backward calculation length and that are consistent with global total emissions”. I think you can only say: “… leads to a posteriori emissions that are less sensitive to LPDM backward calculation lengths than the other baseline estimation methods tested here”

- changed accordingly:

is superior and leads to a posteriori emissions that are far less sensitive to the LPDM backward calculation length and that are consistent with global total emissions → leads to a posteriori emissions that are less sensitive to LPDM backward calculation lengths and stay close to the global total emission value.

L512: "… improves the observational constraint on SF6 emissions substantially". Again, I think we need to know how significant this result is. (Not enough to just demonstrate that the mean has changed in some regions).

- Yes, we agree and changed the sentence accordingly:

The additional use of flask measurements improves the observational constraint on SF_6 emissions substantially. → The additional use of flask measurements has the potential to improve the observational constraint on SF_6 emissions, especially close to the measurement sites.

L517: "Following these results, we strongly recommend to abandon the use of baseline methods based purely on the observations of individual sites, for inverse modelling". I think this statement is too strong. Clearly, studies have shown statistical methods to be consistent with GDB methods for some regions/approaches. My feeling is that you need to be very careful when and where you use them.

- Yes, we agree and changed the sentence accordingly:

Following these results, we strongly recommend to abandon the use of baseline methods based purely on the observations of individual sites, for inverse modelling → Following these results, we advise against the use of baseline methods that are purely based on the observations of individual sites. At least great care needs to be taken that problems such as demonstrated in this paper do not occur.

L519: again, I'm not sure 5-10 days is "usual".

- changed accordingly:

We also recommend to employ longer LPDM backward simulation periods, beyond the usual 5-10 days, as this leads to improvements in overall model performance, helps to constrain emissions in regions poorly covered by the monitoring network, and produces more robust global emission estimates. → We recommend also to employ longer LPDM backward simulation periods, beyond 5-10 days, as this can lead to improvements in overall model performance, can produces more robust global emission estimates and might help to constrain emissions in regions poorly covered by the monitoring network.

L520: "When consistency between regional and global emission estimates is important, even longer backward simulation periods than 50 days may be useful." Again, I don't think you can derive global emissions with very long simulation lengths in the real world. There are other factors that get in the way (low sensitivity, accumulation of errors).

- As mentioned, we think this statement is actually true. Globally, the sensitivity will grow for longer simulation periods and we think that the growing error of individual trajectories will become less and less important (due to the statistical approach of FLEXPART looking at average residence times rather than the individual trajectories). When run long enough (years), FLEXPART produces a well-mixed state of particles (where particle densities are proportional to air density). This is then equivalent to a global box model, and these models have been used for a long time to estimate global emissions. We discuss this in the text (see comment to L12, Final remark).

References

Brunner, D., Arnold, T., Henne, S., Manning, A., Thompson, R. L., Maione, M., O'Doherty, S., and Reimann, S.: Comparison of four inverse modelling systems applied to the estimation of HFC-125, HFC-134a, and SF_6 emissions over Europe, Atmos. Chem. Phys., 17, 10651–10674, https://doi.org/10.5194/acp-17-10651-2017, 2017.

Ganshin, A., Oda, T., Saito, M., Maksyutov, S., Valsala, V., Andres, R. J., Fisher, R. E., Lowry, D., Lukyanov, A., Matsueda, H., Nisbet, E. G., Rigby, M., Sawa, Y., Toumi, R., Tsuboi, K., Varlagin, A., and Zhuravlev, R.: A global coupled Eulerian-Lagrangian model and 1 × 1 km CO2 surface flux dataset for high-resolution atmospheric CO2 transport simulations, Geoscientific Model Development, 5, 231–243, https://doi.org/10.5194/gmd-5-231-2012, 2012.

Hu, L., Andrews, A. E., Thoning, K. W., Sweeney, C., Miller, J. B., Michalak, A. M., Dlugokencky, E., Tans, P. P., Shiga, Y. P., Mountain, M., Nehrkorn, T., Montzka, S. A., McKain, K., Kofler, J., Trudeau, M., Michel, S. E., Biraud, S. C., Fischer, M. L., Worthy, D. E. J., Vaughn, B. H., White, J. W. C., Yadav, V., Basu, S., and van der Velde, I. R.: Enhanced North American carbon uptake associated with El Niño, Sci. Adv., 5, eaaw0076, https://doi.org/10.1126/sciadv.aaw0076, 2019.

Lunt, M. F., Rigby, M., Ganesan, A. L., and Manning, A. J.: Estimation of trace gas fluxes with objectively determined basis functions using reversible-jump Markov chain Monte Carlo, Geoscientific Model Development, 9, 3213–3229, https://doi.org/10.5194/gmd-9-3213-2016, 2016.

Rigby, M., Manning, A. J., and Prinn, R. G.: Inversion of long-lived trace gas emissions using combined Eulerian and Lagrangian chemical transport models, Atmospheric Chemistry and Physics, 11, 9887–9898, https://doi.org/10.5194/acp-11-9887-2011, 2011.

Rigby, M., Park, S., Saito, T., Western, L. M., Redington, A. L., Fang, X., Henne, S., Manning, A. J., Prinn, R. G., Dutton, G. S., Fraser, P. J., Ganesan, A. L., Hall, B. D., Harth, C. M., Kim, J., Kim, K.-R., Krummel, P. B., Lee, T., Li, S., Liang, Q., Lunt, M. F., Montzka, S. A., Mühle, J., O'Doherty, S., Park, M.-K., Reimann, S., Salameh, P. K., Simmonds, P., Tunnicliffe, R. L., Weiss, R. F., Yokouchi, Y., and Young, D.: Increase in CFC-11 emissions from eastern China based on atmospheric observations, Nature, 569, 546–550, https://doi.org/10.1038/s41586-019-1193-4, 2019.

Thompson, R. L., Sasakawa, M., Machida, T., Aalto, T., Worthy, D., Lavric, J. V., Lund Myhre, C., and Stohl, A.: Methane fluxes in the high northern latitudes for 2005–2013 estimated using a Bayesian atmospheric inversion, Atmospheric Chemistry and Physics, 17, 3553–3572, https://doi.org/10.5194/acp-17-3553-2017, 2017

Vollmer, M., Zhou, L., Greally, B., Henne, S., Yao, B., Reimann, S., Stordal, F., Cunnold, D., Zhang, X., Maione, M., et al.: Emissions of ozone-depleting halocarbons from China, Geophysical Research Letters, 36, https://doi.org/10.1029/2009GL038659, 2009