EGUsphere, referee comment RC2
https://doi.org/10.5194/egusphere-2022-275-RC2, 2022

Comment on egusphere-2022-275

Anonymous Referee #2

Referee comment on "A comprehensive evaluation of the use of Lagrangian particle dispersion models for inverse modeling of greenhouse gas emissions" by Martin Vojta et al., EGUsphere, https://doi.org/10.5194/egusphere-2022-275-RC2, 2022

General

The manuscript '' is an important contribution to the field of inverse modelling of synthetic gases. While I largely agree that the use of global concentration fields in combination with backward Lagrangian modelling is superior to the use of observation-derived baselines and should be employed wherever possible, I still feel that the present study is not sufficiently general to arrive at the strong conclusions and recommendations it makes. There certainly are situations and alternative descriptions for which observation-based approaches can yield unbiased emission estimates and, due to the lack of realistic three-dimensional representations of the target compound, present the only reasonable approach. Otherwise the presented study is of high quality, well presented and should be published as soon as the points below were addressed.

Major comments

Very general conclusions: The discussion and conclusions of the present study are not well balanced. They are generalizing the findings of this specific study in a way that seems overreaching. This becomes especially apparent in the conclusion section where very general recommendations are made. From the presented results it is only apparent that certain inversion setups do not work well. However, the current study is incomplete in the sense that only a relatively specific kind of baseline treatment was analyzed in a single global inversion system, while there are several other systems and approaches used by

other research groups and it remains unclear if these other approaches suffer from the same problems. The next points give two alternatives to the solution suggested in the present study. These should at least be considered in the discussion of the manuscript and the conclusions should be amended accordingly.

No consideration of other commonly used baseline methods: The paper focusses on two observation-base baselines methods that have been used in regional inversions of synthetic gases (REBS and Stohl), but it does not mention that other approaches exist (beside the suggested GDB method). Most notably, the method applied in the UK by the MetOffice and Uni Bristol groups seems worth mentioning (see for example Manning et al. (2021) and references therein). While their method (similar to Stohl's method) combines observations and model information, it does not assume that the baseline is a smooth curve, but it depends on the direction and height of the air entering a regional domain. In this way is it possible to describe temporally varying baselines that do not suffer from some of the problems described in the current paper. Admittedly, their method is supposed to be used for regional-scale, limited area inversions. However, it has been used before in a combined Eulerian/Lagrangian study on SF6 (Rigby et al., 2011) and is certainly worth mentioning.

No optimization of baseline: Most regional scale inversion studies that utilize some kind of baseline estimate (if purely observation-based, including transport information, or larger scale concentration fields) don't deny that there may be problems with these assumed/derived baselines. As a consequence, they optimize the baseline in some way in the inversion. Here, the authors chose to apply baselines without further optimization and come to the conclusion that simple baseline approaches should be avoided. However, it would be interesting to see how much an appropriately configured baseline optimization step could remedy some of the problems encountered for example for the REBS and Stohl baselines. Such a test would also be very valuable in the context of biased larger scale concentration fields (see L241, L354), such as used here from FLEXPART-CTM, but more general when derived from satellite data assimilation. I would encourage the authors to do an additional set of inversions where they optimize the REBS and/or Stohl baselines and add such results to Figure 10. In addition, similar tests could be done with the biased baseline fields as presented in Figure 14 c/d.

Minor comments

L25: Not sure that Henne et al. (2016) is the most general reference for promoting inverse modeling methods. Articles like Nisbet and Weiss (2010), Weiss and Prinn (2011) or Leip et al. (2018) have a much broader claim on the subject. Just to mention a few.

L32 and elsewhere: The term 'in-situ' seems to be used to distinguish continuous from flask sampling. In my understanding of the term, both are in-situ, even if the second is not analysed in the field. This is opposed to remote sensing observations. I would suggested to distinguish between continuous and flask sampling. The use of satellite observations for inverse modelling could be mentioned at this point of the manuscript as well.

L34f: Already here, it would be important to distinguish between regional and global scale studies. Most (all) of the cited papers focus on regional scale for which longer integration times are not necessarily useful because particles will have left the regional domain for which emissions are determined!

L49f: I wonder why REBS was selected as a baseline method. It is not the official baseline method operationally applied within the AGAGE network and used for many regional and global inverse modelling studies. While the AGAGE method (also referred to as Georgia Tech method (O'Doherty et al. (2001)) is purely observation-based, methods using additional model information are commonly used within the AGAGE community as well (Manning et al. 2021). Adding some of this information to the introduction would be useful.

L63-65: At this point it is not clear what you mean by space domain. I suppose all of the mentioned studies couple in space (domain boundaries) and time. Even Thompson and Stohl couple in time and space, since the sensitivity field is used, just like in the present study, right? I suggest to rephrase these two sentences and clarify the difference. Another

study to mention here is that by Rigby et al. (2011), which followed a similar approach for SF6, but is not discussed for that reason so far.

L76f: Citing Rigby et al. (2011) would make sense here as well.

L104/105: The observation treatment is a bit unclear. According to this sentence, observations in a four-hour window were selected (12:00 to 16:00) and then aggregated to 3-hourly intervals. Does this imply that two aggregates were formed (if observations were present). For example one for 12:00 to 15:00 and one for 15:00 to 18:00 if local time is UTC? Why the four-hour window instead of simply taking a single three-hour window aligned with the simulations?

L108: If I recall Stohl et al. (2009) correctly, they did not remove observations but assigned larger uncertainties to those where the mismatch was large. Doing so in an iterative step. Does the current approach, hence, differ from Stohl et al. (2009)?

L112: Why was the year 2012 chosen for this study? Wouldn't there be more and more precise observations of SF6 for more recent years?

L113: At this point it is not yet clear what the 're-analysis' is. This only becomes clear in section 2.3. Somehow introduce the re-analysis and motivate as well why two instead of one year of observations were needed (once more only clarified in 2.3).

L143ff: Please mention number of grid cells in inversion grid. Later on (section 2.6), a temporal correlation is mentioned. Please clarify if x varies with time and if so what is the temporal resolution.

Fig2: Does the presented source receptor relationship take the variable sampling frequency at different sites into account? To me it looks as if the flask sampling sites have a very similar source receptor relationship as the continuous sites, whereas in reality they should have total sensitivities at least one order of magnitude smaller (one weekly sampling vs two 3-hour aggregates per day).

L245: If I understand correctly, the same observations are first nudged into FLEXPART-CTM and then used for the global inversion step. While, this seems to be great to remove any biases in baseline concentrations it also means that baseline and inversion are not independent, which may require additional considerations for the Bayesian inference.

L247: What is the rational for using 12 million model particles? Later on, it is mentioned that this may limit the quality of the derived concentration fields. Would a doubling of the particle number have helped improving the concentration fields?

L250: Is a single year spin-up sufficiently long to get the vertical profiles into equilibrium?

L253: What is the temporal resolution of the output? Was it used in this resolution for the coupling to the backward simulations?

Sec 3.1: Although, the situation of the two stations is described, it is not spelled out why these two sites were selected. I assume to show one polluted vs one clean site. Findings for these sites may therefore be extreme. Please discuss this when introducing the sites and please add that Ragged Point, as an equatorial site, is intermittently impacted by southern and northern hemisphere air, which makes baseline estimation a challenging task (as seen later).

L281f: One has to read until the end of the paragraph to get the link to the figures. Would be good to have this from the beginning of the description.

L286f: The sentence is a bit hard to grasp. Consider rephrasing.

L299f: Agreed: but the remedy is to choose a backward integration time that fits the definition of the baseline. The backward integration time is usually chosen such that a released tracer would become well mixed within a latitude band with this period and therefore unobservable as such. This is usually assumed to be around two weeks. 50 days certainly is a bit long and, hence, some of what is background from the observational point of view gets mixed in into the 'recent' signal. It would be interesting to see the concentration increases between day 20 and 50. How variable are they? From Fig 7e/f this seems to be a rather constant contribution. This should be discussed along with the motivation for using smooth baselines.

L314f: But then again the absolute bias is larger than with the REBS method. So it's kind of difficult to say which method is superior here? There is another obvious problem with Stohl's approach for Ragged Point. The lowest concentrations are obviously (as shown later) due to southern hemispheric influences. Stohl's baseline, hence, is more representative for southern hemispheric conditions. However, these do not necessarily dominate at the site. Ragged Point is certainly a fine example where both methods are

predestined to fail, because there is no such thing as a smooth baseline for this site due to the large inter-hemispheric concentration gradient and the intermittent hemispheric influence.

L320: When showing the comparison of the observations to FLEXPART-CTM output (0 day backward), it would be good to mention that one would expect a close agreement, since these observations were used for nudging in FLEXPART-CTM. So no big surprise that they fit so well in the case of Ragged Point.

L329/330: Somehow seems to contradict the later conclusion that long integration times are important.

L356f: But the local FLEXPART-CTM cell should not have an impact for longer integration times. Could there be a bias between the observational data that was used for nudging over North America and the observations at Mace Head. These are two different networks, right? Then there is the possibility that the nudging over North America is not sufficiently strong to remove any bias introduced by North American emissions. Have you checked the CTM performance at the nudging locations? The inversion corrects East Coast emission down (Fig 8). So maybe they drive a baseline shift in the prior.

L365f: In the presented case the GDB method is also not independent form the observations, since these were used for nudging. So it is not surprising that there is only a small bias!

L367: I think the opposite is true. It is really surprising that REBS-based simulations perform so well. A smooth curve fit through random data would not result in large

correlations only in no bias. The main reason for the good correlation here is that there is a trend in the time series, which is considerably large compared to the pollution events. REBS captures this trend very well. In contrast, GDB may contain a fair fraction of noise (as mentioned elsewhere) and even if the trend is correctly captured, this will lead to lower correlation.

L373: But overall, the 10 d REBS has best correlation and MSE and bias are only slightly worse than for GDB at 50 days. It's a bit difficult to see the large benefits of the GDB method just from the statistics presented in Table 2.

Figure 9: Consider using different y-scale for each country. Adding uncertainty estimates would be valuable as well.

L385: How can we conclude that the increments estimated with the REBS baseline are wrong? Based on the assumption that GDB is correct? Maybe simply formulate in a more careful way as done a few lines below for the Stohl baseline case.

L390: Overall I would mention that patterns in 8b and 8d are more similar than between 8c and 8d, with the exception of East Asia.

Figure 10: Consider adding posterior uncertainties.

Figure 13: Consider adding information on backward integration time to the figure caption.

L336ff: The conclusion that longer integration times allow for correct emission estimation in 'under-sampled' regions should be drawn more carefully. Yes ,technically this is true. However, the problem remains that the sensitivity to these regions (given the current network) is small compared to regions where the observations are taken. Small sensitivities may simply lead to more random adjustments in the inversion, because they would only contribute small changes to the observation mismatch. An observing system simulation experiment would be better suited to prove this point than analyzing uncertainty reductions and assuming that the 50 d inversion does the job correctly. Furthermore, one may get the idea that with the 50 d (or longer) integration we do not need any more flask sampling sites, something that is afterwards promoted in the manuscript.

L447: The baseline could also be taken from nearby or same latitude continuous sites or, as in the method by Manning et al. (2021), could be represented through baselines at the domain border (not possible for global runs).

L450f: What is the impact of the additional flask data on the national and global emission estimates. Maybe add to Fig 10 and 11.

L469ff and Fig 14c-f: I cannot follow the suggestion of how the bias in baseline is supposed to lead to the large bias in the global total emissions as given by the pink lines in Fig14. If I understand it correctly, the expectation is that the inflicted baseline bias would need to be compensated by increased emissions during the period of backward integration. Somehow, there seems to be some misconception here. Although, this consideration would make sense for a global (box) model that is run with baseline concentrations to estimate global emissions, it cannot be applied to the kind of observations and "regional" simulations done here. The sampled concentration peaks do not represent fully mixed emissions, but recent emission impacts. The bias of 0.003 ppt is orders of magnitude smaller than the regional emission signal simulated at the observation sites, even if only 1 day backward transport is considered (see Fig 4 and Fig

5). However, it is this regional signal that is used in the inversion step, not the annual global trend. Offsets in the regional signal in the order of 1 % will hardly have an effect on the emissions that is in the order suggested by the pink line in Fig 14. I would suggest to redo the sensitivity test with a biased background using a much larger bias then suggested here. How large the magnitude of this bias should be is hard to tell, but maybe it could be taken as the difference between REBS and Stohl baselines. Alternatively, a different interpretation and a re-thinking of the expectation (pink line) would also help these sensitivity tests.

L492ff: Consider 'may lead' instead of 'leads to'. REBS baselines can still work as you show in Figure 10. It would be fair to mention here that the performance for integration periods that are typically used 10 or 20 days are in better agreement. Also the fact that most inversion systems try to optimize baseline biases as part of the inversion, should be mentioned.

L497ff: Same as above: mention that baseline biases could be treated as part of the inversion.

L505ff: From Table 2 I can only conclude this for the GDB method. The other two methods show insignificant improvements from 10 to 50 days or even worse performance (in terms of bias). I would also think that this improved performance from 10 to 50 days for GDB would strongly depend on the baseline model. Here, a relatively coarse model is used. Higher resolution may result in very good performance already at shorter backward integration times.

L508: Again, I find this statement too general. The problem with the biased prior results from the fact that with short integration times there is no sensitivity to large areas, so there is no chance for the inversion to correct this. If observations would cover all emitting areas well within 10 days the bias should also be removed. Many regional and global scale inversions result exist where posterior emissions moved far away from the prior, but the key is observational constraint. If there is little constraint on certain elements of the state vector, we cannot expect the posterior result to be more accurate than the prior.

L515ff: This is very much in line with what Weiss et al. (2021) suggested as well. Why not mention that?

L517ff: Both statements are very general. For regional inversions longer integration times don't necessarily make more sense. Baselines can be sampled from conditions at domain border, either from global model as in GDB or constructed from observations (like Manning et al., 2021). Similarily, the optimization of the baseline as part of the inversion to avoid biases, should be mentioned. This will still be necessary for the GDB approach when biased global fields are used.

Technical corrections

Equation 4: The last term should contain lower case x for the prior state vector.

Additional references

Leip, A., Skiba, U., Vermeulen, A., and Thompson, R. L.: A complete rethink is needed on how greenhouse gas emissions are quantified for national reporting, Atmos. Environ., 174, 237-240, doi: 10.1016/j.atmosenv.2017.12.006, 2018.

Manning, A. J., Redington, A. L., Say, D., O'Doherty, S., Young, D., Simmonds, P. G., Vollmer, M. K., Mühle, J., Arduini, J., Spain, G., Wisher, A., Maione, M., Schuck, T. J., Stanley, K., Reimann, S., Engel, A., Krummel, P. B., Fraser, P. J., Harth, C. M., Salameh, P. K., Weiss, R. F., Gluckman, R., Brown, P. N., Watterson, J. D., and Arnold, T.: Evidence of a recent decline in UK emissions of hydrofluorocarbons determined by the InTEM inverse model and atmospheric measurements, Atmos. Chem. Phys., 21, 12739-12755, doi: 10.5194/acp-21-12739-2021, 2021.

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

O'Doherty, S., Simmonds, P. G., Cunnold, D. M., Wang, H. J., Sturrock, G. A., Fraser, P. J., Ryall, D., Derwent, R. G., Weiss, R. F., Salameh, P., Miller, B. R., and Prinn, R. G.: In situ chloroform measurements at Advanced Global Atmospheric Gases Experiment atmospheric research stations from 1994 to 1998, J. Geophys. Res., 106, 20429-20444, doi: 10.1029/2000JD900792, 2001.

Weiss, R. F., and Prinn, R. G.: Quantifying greenhouse-gas emissions from atmospheric measurements: a critical reality check for climate legislation, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 369, 1925-1942, doi: 10.1098/rsta.2011.0006, 2011.