Response to review comments

Referee 1

This paper outlines the development of an inverse modelling software package "FLEXINVERT". The methods employed have mostly been published elsewhere and are well known. Therefore, whilst there are relatively few new insights in this article, the paper is a very thorough and clear account of FLEXINVERT system. I think it will be suitable for publication in GMD, once the following comments have been addressed.

We thank Referee 1 for his/her constructive comments and reply to them below.

General comments

One of the main findings in the paper is how sensitive certain parts of the inversion are to "background" mixing ratios. This is well known for Lagrangian model inversions, but I think it merits further discussion here. In this paper, baselines are either estimated using an Eulerian model, or from the mole fractions directly. There are number of problems that occur to me: a) In the case that optimized TM5 mole fractions were used, there is an element of circularity, because the TM5 mole fractions will have already seen the observations. b) In the case where the lower quartile of observations were used, did the authors take into account the fact that the lowest measured mole fractions can have a wide range of origins? In particular, for methane, when air enters Europe from the Southerly sector, it can be significantly depleted in methane, compared to other "baselines" from the Atlantic? If I understand their method correctly, the baselines they obtain would be rather smoothed, and would not identify short-timescale "low methane" events. Even more significant "depletions" have been observed elsewhere in the world, for example in East Asia, where the air can rapidly fluctuate between Northern and Southern hemispheric during the summer. c) In the case where these baseline mole fractions were optimized, there is also an element of circularity, as the observations themselves have been used to determine the "prior". In each of these cases, there is the potential for the choice of baseline to erroneously influence the derived emissions either through biases (that would likely not be well accounted for in the uncertainty quantification method outlined, which assumes only stochastic errors). I don't think the paper needs to solve these problems. However, I think the discussion could be expanded very slightly to further highlight some limitations.

The problem of accurately determining background mixing ratios for Lagrangian models is a challenging one. We address the two cases the referee mentions as follows: 1) using the optimized TM5 mole fractions. If we consider just using the background directly from the TM5 optimized fields, i.e. not including these in the optimization of the state vector in the Lagrangian inversion, then there is little circularity, but rather this can be seen as a two-step optimization. In this case, the observations are used to optimize the background in the first step and the fluxes in the second step. On the other hand, if the background variables are also included in the state vector, then there is a degree of circularity. However, the two inversions, i.e. the TM5 and the Lagrangian inversion, do not use the exactly same set of observations nor at the same time frequency, uncertainty estimate or transport, therefore there is

still new information when optimizing the background in the Lagrangian inversion. Generally, global models, such as TM5 are optimized using a standard observation dataset based on e.g. the NOAA GMD network, while a regional inversion (in this case the Lagrangian inversion) makes use of other observations instead of or in addition to those used in the global model thus minimizing or avoiding the problem of circularity. It should also be noted that the background calculated for the Lagrangian model is an extremely smoothed version of the Eulerian mixing ratio fields, as the sensitivity to the background even for a single measurement is distributed over large parts of at least one hemisphere. In the global model, this background is constrained mainly by measurements from outside the region of interest. In contrast, the influence of the stations used for the regional inversion (which typically are concentrated in the area of interest) on the background as sampled by the Lagrangian model is quite small. This minimizes the degree of circularity inherent to the use of optimized concentration fields.

2) using the lower quartile of observations. We tested a number of methods for calculating the baseline using the observations themselves, including also higher temporally resolved baselines. We found from these tests that having step jumps in the baseline can cause problems in the optimization of the fluxes given the imperfect transport and often weak observational constraint. Instead, we found a smoothed baseline to be more robust although this does not take into account instances where the background has considerably different mixing ratios, which as the reviewer states, is an important problem and was also mentioned by Stohl et al. (2009). In their study, Australian emissions were overestimated, as they were constrained mainly by measurements from a single station (Cape Grim), which is located in the south of the emission area and for which a background was derived that is mainly representative for air masses coming from the west. The derived background, however, was systematically too low for air flows that are coming from the north (where the Australian emissions are located). This was compensated in their inversion with unrealistically high emissions in Australia. Such problems are much less severe for emission regions surrounded by several measurement stations, for which background biases will be in the opposite direction and will partly cancel each other. Future work, could involve combining information from the Lagrangian model about where the air masses originate (especially the latitude) to help better define the observation-based background. In any case, the degree of smoothing, i.e. the window length over which to calculate the lower quartiles and the averaging is a parameter that can be changed to best represent the variability in the background. Concerning the circularity of using the observations to determine the prior background, the observations that correspond to background are those that have very little influence from fluxes within the domain. Thus the corresponding rows of the transport operator, H (or the Source-Receptor Relationships = SSRs), are close to zero, thus these observations will not contribute much to the constraint of the fluxes within the domain.

We have now expanded our discussion in section 3.4 to include both the concern of circularity and the temporal resolution of the background calculated from the observations.

Specific comments

Page 3754, line 13: I'm not sure what this sentence means, or whether "smearing" is

the best word to use.

We mean that a Lagrangian model can have infinitesimal resolution and thus is able to better resolve measurements at a site while an Eulerian model has fixed resolution and the measurement site is represented by the value of the corresponding grid cell (or in some cases the interpolation between grid cells). We have changed this sentence to:

"A further advantage of LPDMs is that they can be run backward exactly from a measurement site, unlike Eulerian models, where site measurements are represented by the averaged value of the corresponding grid cell."

Page 3761, line 11 and Equation 6: I'm not entirely clear why this is necessarily an aggregation error. I think you can formulate this problem so that y_mod for the variable grid is identical to that obtained using the full grid, by using the emissions-weighted footprint. In that case, the aggregation error would only come in during the inversion.
It should be noted that H_nest_vg*f_nest_vg is not equivalent to H_nest*f_nest. The difference between these is the aggregation error.

Equation 7: I think F_out should be lower case

In Eq. 7 we have used F_out as this is a matrix consisting of row vectors of f_out. Thus F_out and f_out are not equivalent.

Page 3762, Line 26: I think this should be a numeric "1", rather than "one"

We have changed "one" to "1".

Page 3766, Line 18: This idea has been used elsewhere. Perhaps a reference or two should be given.

We have included a reference to Bergamaschi et al. 2009.

Page 3768, Line 2: superscript "T" for transpose in the line below the equation.

This has been corrected.

Section 2.9: If I read this correctly, it appears that this "non-negativity" correction only applies to those grid cells where negative emissions were obtained? I reality, if the first inversion could "see" this constraint, wouldn't its effects be felt further away than the individual grid cells where negative emissions were derived? Perhaps a line or two of clarification, could be provided.

In fact, the non-negative constraint (given in Eq. 20) does not only affect the grid cells with negative values but there is also some adjustment to other cells according to the correlations described by the posterior error covariance matrix, A_flux. This is now also stated in Section 2.9:

"The inequality constraint does not only affect the grid cells with negative values but there is also some adjustment to other cells according to the correlations described by the posterior error covariance matrix, $\mathbf{A}^{\mathbf{flux}}$."

Page 3768, Line 26: Appendix C

We actually refer to the description of the full code and not just to the inequality constraints. This sentence was supposed to be a new paragraph. Appendix B is the correct reference.

Page 3770, Line 22: What about the representation error at the flask sampling sites?

Since flask measurements were assimilated without averaging (and by choosing the closest modelled back trajectory to the flask sampling time) no representation error was included for these, i.e. only the model and measurement errors were used.

Page 3774, Line 24: This assertion could be tested by running the model at a higher release height.

We tested this by running the Lagrangian model, FLEXPART, at a height of 2000 masl to represent PUY. The correlation coefficient and normalized standard deviation for the observations and the simulation using the prior emissions with 2000 masl was 0.25 and 2.09, respectively, compared with 0.18 and 2.24 with 1475 masl, i.e., the actual height of PUY. It is likely that at 1475 m, the influence of local emissions at PUY is overestimated owing to a too deep BL in FLEXPART, meaning that PUY is too often in BL compared to the reality.

Referee 2

General Comment

I think that the work and method described are good. It could be a very useful tool for people to use in understanding emissions. The key point is that many assumptions are required to be made and the user needs to understand what they are before this tool is used.

We thank Referee 2 for his/her constructive comments and reply to them below.

Specific Comments

p.3756 l.4: This is only true if the original LPDM run is on a fine enough resolution and large enough domain to meet all needs and also not if applied to gases with different loss processes - unless all particle info retained.

We have now modified this sentence to:
" the LPDM needs only be run once for each species and receptor to find the SRRs, as the output can be applied to optimize the fluxes for any domain and resolution (as long as the resolution is no finer than that of the LPDM run)."

p.3757 Eq2: I do not like the notation Hbg. It has different units to Hnest and Hout.

While it is true the H_bg has different units to H_nest and H_out, we prefer this notation as "H" indicates in general that this is a transport operator.

p.3758 Eq3: time/density has units time x volume / mass not time/volume as stated. What am I missing? Please clarify.

This was an error, it should read: "SRR is in units of residence time times volume per mass". We have corrected this now.

p.3759 l.7: I assume that ni trajectories terminate in the grid cell? Please re-word to make clearer.

n_i is the number of particles that terminate in a given grid cell (or in other words the number of particle trajectories that terminate in a given grid cell). We have changed this sentence to:
"The sensitivity to mixing ratio in a grid cell at a given time (i) is calculated as the number of particle trajectories that terminate in the grid cell (n_i) divided by the total number of particle trajectories released (J)".

p.3759 l.22: Only works if site 'sees' background air 25% of the time. This part ignores the influence of latitude / altitude that is actually known. The next part also assumes that the prior is correct and the site not too polluted. It was unclear to me how the timeseries from lower quartile obs - (prior x SRR) was combined with just the time-series of (prior x SRR)? Are these steps sequential or exclusive?

Like any method that tries to extract the background from the observations (for a discussion, see Giostra et al., 2011), this assumes that the background signal is actually observed on a semi regular basis at the measurement site. This is, however, not always the case as the reviewer points out. We chose to select the lower quartile of values, which assumes that background is observed at least 25% of the time. The percentile of data selected for the background calculation can be varied to best represent the data. This method does not ignore the variation in the background with changing latitude/altitude of origin of the air masses as the change in mixing ratio with latitude/altitude will be reflected in the variability of the time series. A greater problem is that of resolving the variability in the background, e.g. on synoptic scales, which requires in situ measurements, rather than discrete ones, and selecting an appropriate window length over which to select the lower percentile and for the smoothing.

Once observations have been selected as being "background" (in our study we used the lower quartile), then the prior modelled values are calculated for these times. If the observation is truly background, there should be little to no influence from fluxes within the domain, thus the prior modelled value should be small. We subtracted these prior values from the "background" observations (thus the steps are sequential). In some cases, the prior value was higher than what would be expected for a background observation, in these cases, we subtracted a smoothed prior value instead. Although, this does imply that the prior is correct, the error in doing this is likely only small since the prior values themselves are small.

Determining the background reliably is very challenging, and requires assumptions somewhere about the validity of the prior model and what the observations represent.

p.3761 l.2: Helpful to add a comment that =0 as each fine grid can only be in one coarse grid.

We have added this now.

p.3762 l.2: Change 'row' to 'rows'.

We have corrected this.

p.3762 l.11: I assume that a measurement can have contributions from several latitudinal bands rather than just one as stated here, it is contradicted in the next paragraph, please clarify and re-word.

In the current set-up for the case using the observation-based method for determining the background, we do not disaggregate the background contribution to separate latitudinal bands as there is no information to constrain what the contribution from each latitudinal band is. For the case using the model-based method for determining the background, the contribution is disaggregated over latitudinal band (and optionally also by longitude) so that different latitudes (and longitudes) contribute to the background for each measurement. We have now changed the text to make this clearer.

P.3761-2, Eq 5 and 9: I would suggest the use of the word 'and' rather than commas

We have replaced the commas with "and".

p.3763 l.8: Please either use 'modelled' or 'modeled' throughout or whatever the journal requires, both are used at the moment.

We have changed this now to "modelling" and "modelled" in accordance with British spelling (except in the bibliography where we use the original spelling).

p.3763 l.8: It does not need to stay within the bounds does it? Are they not expressed as standard deviation uncertainties and therefore have tails? What bounds are being described here?

The solution does not have to stay within any bounds, what we meant was only that the most probable solution depends also on the probability of the prior, which is described by its mean value and standard deviation uncertainties. We have now changed this sentence to:
"Based on Bayes' theorem, the most probable solution for $\mathbf{x}$ is the one that minimises the difference between the observed and modelled mixing ratios while also depending on the prior state variables, $\mathbf{x_b}$ and their uncertainties"

p.3763 l.21: Out of interest, why if M»N is it easier to invert MxM matrix? I am obviously missing something but it struck me as odd.

We apologize, the sentence should actually read: “if the number of observations is smaller than the number of unknowns”. We have corrected this now.

p.3763 l.24: ’The’ rather than ’This’

We have changed this.

p.3764 l.20: What happens if the uncertainty of the prior flux is provided? Is this method still imposed?

We have described the code as it currently is. The prior flux uncertainty may be provided instead, which would involve only a minor change to the code, basically to read this information in and replace the default uncertainties with those provided. However, unfortunately, most emission inventories nowadays still do not provide proper uncertainty values that could be used.

p.3766 l.8: I suspect that the errors in the meteorology e.g. Boundary Layer etc, will be much larger than stochastic uncertainty. So uncertainty will be under-estimated. Obviously not much can be done but a sentence describing this would be helpful.

Yes, we certainly agree. The stochastic uncertainty is likely to be small relative to the transport errors (from PBL height uncertainties etc). We have included the following sentence pointing this out:
“Therefore, we do not quantify the full transport error, but only the part of it that can be estimated from the model FLEXPART, i.e. the stochastic uncertainty, which arises by the representation of transport with a limited number of particles (see Stohl et al. (2005)). The stochastic error, however, is likely to be much smaller than the full transport error.”

p.3766 l.20: Is the ’minimum error’ user defined, please state.

Yes, the minimum error is user-defined. We have now made this clearer in the text.

p.3767 Sect 2.8: Not sure this is a good idea to retain within the paper. I would suggest just removing this section, it seems rather unfinished work.

We possibly over-stated the uncertainty about this part of the code. We have tested this functionality in a limited number of cases and found that it works. However, we added the caution message, namely that the output should be evaluated carefully, because we cannot be sure of how stable this is in all cases. We have removed “still experimental”.

p.3768 Sect 2.9: ’error-free observations’ - this needs more discussion. What does it mean for the results? What assumptions are being made? Are the uncertainties being affected?

We have added the following sentences to section 2.9:
“The inequality constraint does not only affect the grid cells with negative values but there is also some adjustment to other cells according to the correlations described by the posterior error covariance matrix, $\mathbf{A}^{\mathbf{flux}}$. The posterior error covariance matrix,

however, is unchanged since the observation error covariance matrix in this case is zero."

p.3769 l.5: Why is the reaction with OH a good reason to choose CH4? Was it because it has a linear loss process?

We meant rather that because it is principally lost by OH reaction, which is approximately a linear process, that it is a suitable species to use for the test. We have reworded this as follows:
"Methane was chosen, as it is an important greenhouse gas with an atmospheric lifetime of approximately 10 years (Denman et al. 2007) and since its loss in the troposphere is principally by reaction with the OH radical, which can be approximated as a linear process."

p.3769 l.12: Suggest replacing 'and largely wetlands and' with 'principally wetlands'

Done.

p.3769 l.25: Suggest replacing 'quite' with 'very'

Done.

p.3768 Sect 3.1.1: No mention of release height for mountain stations, agl or asl and a sentence to describe the issue here.
We released the particles from the approximate height of the sampling inlet i.e. the station height above sea level plus the height of the inlet above ground level. For mountain stations, we used the given altitudes. We have added the following sentence:
"Particles were released from the sampling inlet height at each observation site (see Table 3)"
Also, we specify in Table 3 that this is the height in meters above sea level.

p.3769: It would appear that LMP and CIB have the same impact as in-situ sites, were the different number of observations per site not used in calculating Fig 2a? I suggest this is important.

Fig. 2a shows the footprint calculated using FLEXPART assuming the same number of observations at each site. This is the information that is used to determine the variable grid, i.e. the different number of observations at each site is not taken into account in this step. We have now added this information to the caption of Fig. 2a.

p.3770 l.28: Add except high altitude sites to be consistent with p.3774 l.3.

We have added this.

p.3770 l.28: Boundary layer averages?

Yes.

p.3770 l.20: 'closest available' - I assume you mean 3-hr map that encompasses the observation?

We mean the SRR (or "footprint") corresponding to the closest 3-hourly particle ensemble release (we use the terminology "retro-plume" to distinguish from single trajectories). We have now tried to make this clearer in the text.

p.3770 l.22: It would appear that the in-situ obs are given greater uncertainty as they also have a representation error (variability error maybe better term). In fact the opposite is true, the flask data is only one point so nothing is known about the variability. This therefore should be higher than the in-situ data.

Our reasoning behind using the SD of the in-situ observations over the averaging interval for the representation error is that these are not repeat measurements of the same thing, therefore, the uncertainty does not decrease with averaging the observations. Instead, we are losing some information in averaging, so in this respect, it may also be thought of as a temporal aggregation error. On the other hand, we do not average the discrete (flask) measurements so there is no associated aggregation error.

p.3771 l.17: Therefore some observations are used twice - please state this and comment on the possible effect.

It is true that some observations may be used twice, i.e. first in the optimization of TM5, which is used to determine the background, and second, in the optimization of the fluxes on the nested domain. However, new observations were used in the Lagrangian inversion that were not used in the optimization of TM5 and the averaging interval at which they were used differs. We have added a discussion of this problem to section 3.4 (also in response to a comment by Reviewer 1).

p.3771 l.25: It was interesting to note that uncertainty in background was though more certain than measurement error - I would think the opposite?

The 0.2% refers to the uncertainty in the scalars of the background, which, if the background is optimized in the inversion, are included in the state vector. For background values of circa 2000 ppb this uncertainty is close to the measurement uncertainty. The concern was that increasing the uncertainty in the background scalars, would allow too many degrees of freedom for adjusting the background and, as our tests showed, can lead to erroneous background values a posteriori.

p.3772 l.25: Highlights that considerable care is required. Is this an issue with the prior or poor modelling of MHD? Prior has a dominate effect and the uncertainty in the prior is not propagated.

The determination of the background at MHD from the observations is complicated by the fact that the signal from within the domain is very small, i.e. the observations are mostly background. The overestimate of influence of emissions in the domain can be due to both errors in the transport and errors in the prior emissions. It is not possible to disentangle these two effects. The uncertainty in the prior may be propagated into the uncertainty in the background.

p.3773 l.11: Suggest removing 'slightly'

We have removed "slightly".

p.3774 l.25: Link this to discussion on other mountain sites used in the inversion JFJ and CMN etc as mentioned earlier.

JFJ and CMN are, compared to PUY, quite well modelled by FLEXPART and the normalized standard deviation at both these sites is very close to one, while it is close to 2 at PUY. We think, therefore, that PUY is not well modelled owing to both the topography (the station is located on a volcanic cone, which represents a very abrupt change in topography) as well as the fact that there are significant emissions in the prior around the station. A likely explanation is that FLEXPART overestimates the BL height at PUY and thus overestimates the influence of local emissions. We have amended the text accordingly.

p.3775 l.15: Relevant observations e.g. Cabauw used in Bergamaschi work which would have a dominate effect in Benelux.

Yes we agree, but unfortunately we could not get access to the observations from Cabauw. We have now included this fact in the discussion as well.

I could not find some of the references that are listed in the actual paper, probably missed some e.g. Etiope, Houweling, Lambert and Sanderson.

Etiope et al. (2008); Lambert and Schmidt (1993) and Houweling et al. (1999) are referenced in Table 4.

Referee 3

We thank Referee 3 for his/her constructive comments and reply to them below.

The manuscript describes an inverse emission estimation framework "FLEXINVERT" for an analytical Bayesian inversion tailored for backward Lagrangian transport models like FLEXPART. Although most of the elements are well known and have been described elsewhere, there are two main reasons why I recommend publication of the manuscript: First of all, FLEXINVERT is presented as a comprehensive framework that combines essentially all of the elements required for such a system including a proper consideration of background concentrations (necessary because the backward transport simulation accounts only for fluxes during the recent history of an air parcel) , the computation of a variable resolution grid reflecting the true sensitivities of the observation network to upstream fluxes, the definition of observation errors and a priori uncertainties (including their spatial and temporal correlations), and the mathematical and numerical framework to solve for the optimized fluxes and their posterior uncertainties. I particularly appreciate the mathematical rigor the individual components are described with. This will greatly help any user of the framework to understand the individual steps.

The second reason why the manuscript deserves publication is that there are innovative elements that (to my knowledge) have not been presented before, at least not in the same way or to the same level of detail: These include the "Aggregation of background mixing ratios" in section 2.3 and the "Optimization of the fluxes to fine resolution" in section 2.8.

The paper is very well written and structured and well to the point. I thus have no specific recommendations regarding structure or content.

Nevertheless, I have one main concern which deserves more attention by the authors, in particular regarding future users that should also know the potential limitations of the framework: What are the computational costs and what are the corresponding limitations? 4DVAR systems have been developed because inverse problems can easily become too large to be solved analytically. FLEXINVERT is based on an analytical Bayesian inversion which involves operations with large matrices (including matrix inverse) that may become computationally expensive in particular in terms of required memory space. This will necessarily limit the applicability to small- to medium-size problems (limited number of observations/measurements sites, limited spatial and/or temporal resolution of the fluxes). Many operations involve sparse matrices but only standard linear algebra methods appear to be applied which will necessarily lead to many unnecessary computations. The observation error covariance matrix was chosen to be a diagonal matrix. Was this choice driven also by computational constraints? Although the treatment of background concentration looks appealing at first sight, it may involve considerable computational cost: The matrix H_bg may be very (excessively?) large since it is dimensioned M (number of observations) x P (number of grid cells of a global model). It would be useful if the authors could provide some information on the memory and computation time requirements e.g. for the CH4 test case presented in Section 3. I found it difficult to judge which of the individual steps is particularly expensive and may therefore require special attention by a user.
Computational cost: we have now added the computational costs of running FLEXINVERT (memory usage and total computation time) for the test S1 (which is representative of all tests) to section 3.3:

Max. memory usage: 18 GB
Mean memory usage: 6.4 GB
Computation time (min): 1.8 days

FLEXINVERT is still being developed and we will develop a conjugate gradient version as well in order to solve larger problems (larger number of unknown variables and observations). For this, it will also be useful to have the analytical version to compare with as an accurate reference solution, which should be closely reproduced by numerical methods.

Standard linear algebra: in fact the matrices are not that sparse. The sparseness of the transport operator, H, depends on how long backwards in time the trajectories are run and the sparseness of the error covariance matrices depends on the correlation scale length used. Subsequent to submitting this paper, we have implemented the use of eigen-decomposition of the error covariance matrices, B_flux_naw, B_flux, and the spatial error covariance of B_flux_vg to avoid storing these in memory. Also, the full

spatio-temporal error covariance matrix, B_flux_vg is not formed directly but rather calculations using this matrix are made using the eigen-values and vectors of the spatial error covariance and the temporal error covariance matrix.

Observation error covariance matrix: this matrix, R, is in fact not diagonal since the aggregation error covariance matrix (in the observation space) is not diagonal. However, we did not account for error correlations between observations. As the reviewer states, this is not a problem for the infrequent discrete observations (flasks) but could be a problem for the in-situ observations. In the case study presented, we averaged the in-situ observations (daily day time average) and thus reduce the correlation between assimilated observations. If the user specifies the correlation of observations, however, this could be easily added to FLEXINVERT.

Background concentration matrix H_bg: this matrix is only stored in memory for one month at a time. We did not find the size of H_bg to be limiting in our case study.

FLEXINVERT makes a number of assumptions (which are often made), and a user will have to be aware that these may not necessarily be met:

Errors are assumed to be Gaussian. As shown e.g. by Stohl et al. (2010), measurement – model residuals may be highly skewed. FLEXINVERT may therefore be sensitive to extremes.

In the case study presented, we did not find the measurement-model residuals to be highly skewed. However, we agree that in such cases, i.e., with highly skewed distributions, that the inversion will be sensitive to extremes (as with any inversion framework). We have added the following text to section 2.7 pointing out the need to be careful about how to treat outlying observations and refer to the work of Stohl et al. (2009) (please note that the correct reference is 2009 and not 2010).

"Another assumption that is made is that the observed – modelled mixing ratio residuals have a Gaussian distribution (Eq. 10 is based on this assumption). Therefore, in cases where the distribution is highly skewed, observations corresponding to the tail of the distribution will have a strong influence on the result of the inversion. FLEXINVERT does not include any component to deal with skewed distributions; however, the influence of observations in the tail of the distribution may be reduced by increasing their uncertainty. For more details about dealing with skewed distributions we refer the reader to Stohl et al. (2009)."

Observation errors are assumed to be uncorrelated in time: This is likely a good assumption for weekly flask samples, but it would certainly not be a good assumption for e.g. hourly data, mainly due to correlated transport errors (not only correlated PBL errors but also errors in the wind field). This assumption can be easily verified by analyzing the autocorrelation structure of the residuals. In this way the correlation length can be determined.

We agree that for e.g. hourly observations, it would be necessary to account for the error correlation between observations and we now point this out in section 2.7. In the case study we present, we use daily daytime averages of the in-situ observations,

therefore, the error correlation between assimilated observations is less than for the hourly data.

The observation-based estimation of background concentrations as presented in Section 2.1.2 will not work for CO2 which has strong negative fluxes and therefore has no clearly defined baseline. It would be good if the authors could add some word of caution on these points.

We certainly agree. The observation-based method for determining the background was included rather for anthropogenic species (e.g. halocarbons) for which there is a definable baseline. We have now added this information to section 3.4.

Finally, results of an inversion critically depend on the specification of a priori and model-data mis-match errors (and their correlation structure). FLEXINVERT provides a nice framework for solving the problem, but it provides little guidance with respect to the specification of these errors. The authors are obviously aware of the necessity to provide realistic error estimates as they have checked their inversion in the case study (Sect. 3) for the chi-square statistics but it should probably be stated more clearly that it is the task of the user to define these errors in a realistic way. A recent publication addressing this issue is Berchet et al. (ACP, 13, 7115–7132, 2013, doi:10.5194/acp-13-7115-2013).

We agree that this is a very important consideration. FLEXINVERT includes a simple scheme for estimating the prior error covariance matrix (described in section 2.5). The prior flux errors are calculated as proportional to the magnitude of the flux with some spatial smoothing, i.e. grid cells with small fluxes but which are adjacent to grid cells with large fluxes will also have a large error. The user can then define the spatial and temporal correlation scale lengths for the describing the error correlations. The total error covariance matrix is then scaled to be consistent with some user-defined estimate of the total error for the domain. We considered this simple scheme to be the best default as it can be applied to many different species. However, the user may decide to use another, maybe more sophisticated scheme.

Minor points:

P3753, lines 14-16: Shouldn't it be "the partial derivative of the change in mixing ratio to the change in fluxes" rather than the reverse?

We refer to the adjoint, thus it is the change in flux that is calculated.

P3754, line 5: Other regional scale inverse modeling studies involving Lagrangian models would be: Keller, C. A., M. Hill, M. K. Vollmer, S. Henne, D. Brunner, S. Reimann, S. O'Doherty, J. Arduini, M. Maione, Z. Ferenczi, L. Haszpra, A. J. Manning and T. Peter, European Emissions of Halogenated Greenhouse Gases Inferred from Atmospheric Measurements, Environ. Sci. Technol., 46, 217-225, doi:10.1021/es202453j ; Brunner, D., S. Henne, C. A. Keller, S. Reimann, M. K. Vollmer, S. O'Doherty, and M. Maione: An extended Kalman-filter for regional scale inverse emission estimation, Atmos. Chem. Phys., 12, 3455-3478, doi:10.5194/acp-12-3455-2012, 2012.

We have added these references to the list.

Page 3756, line 15: Variable resolution grids adapted to the average residence times probably have been introduced for the first time in Manning et al. (JGR, doi:10.1029/2002JD002312, 2003) and have also been applied in other studies such as Vollmer et al. (GRL, doi:10.1029/2009GL038659, 2009), Manning et al. (JGR, doi:10.1029/2010JD014763, 2011), etc. This should not be called the "method of Stohl et al.".

We have changed "method" to "studies" and include the reference to Manning et al. 2003.

Page 3757, lines 15ff: It should be described more clearly that a typical setup of FLEXPART involves an outer (potentially global) domain and a finer, nested domain. The transport of particles is continued in the outer domain once they leave the nested domain. This setup is not always applicable. Consider e.g. a regional scale model such as FLEXPART-WRF where particles may terminate at the borders rather than being transported further. In this case, termination of particles may occur at any time before the end of the simulation. FLEXINVERT does not seem to be prepared for such a case.

The version of FLEXINVERT presented does require that a global domain is used for the FLEXPART runs. To use FLEXPART-WRF, i.e. with a regional domain would require some modifications to account for the fact that the particles terminate at the domain boundary. We have now added this information to section 1:

"FLEXINVERT, as it is presented here, requires that the LPDM is run on a global domain, or at least that the domain is large enough so that trajectories do not exit the domain."

Page 3760, line 8: Estimating background concentrations from observations is a long-standing problem that has been addressed in numerous studies prior to Stohl et al. (2010) and more sophisticated methods have been developed than presented here, e.g. Thoning et al., (JGR 94, 8549–8565, 1989), Ruckstuhl et al. (AMT, doi:10.5194/amt-5-2613-2012, 2012, 2012).

We agree that this is a long-standing problem, and it is not one that we have aimed to address in our paper. Rather, we have adopted a simple routine to calculate the baseline that we found to be fairly robust to the number of observations (i.e. can be used for in-situ as well as discrete measurements). We consider the method of Thoning et al. on the other hand not to be a suitable alternative. Thoning et al. use a fourier transform to filter high frequency (synoptic and shorter timescales) components of the signal, however, their technique requires gap-filling of the data, which may introduce errors. Furthermore, it may in some cases overestimate the baseline if the baseline is at times lower than the low frequency signal (e.g. seasonal cycle) as it could be filtered out. On the other hand, the method of Ruckstuhl et al., could be an interesting alternative, as their method does not require any gap-filling and could be applied to data with differing sample frequency. We have included the following statement in any case to point out that there are a number of alternative methods that could be used:

"This method was chosen as it is robust to the number of observations (i.e. it can be used for in situ as well as discrete measurements) although other more sophisticated background selection algorithms exist (e.g. Ruckstuhl et al. 2001, Giostra et al., 2011)."

Page 3763, lines 19-21: I don't understand why this is the most efficient method when the number of observations M is large compared to the dimension of the state vector. The matrix to be inverted has dimension M x M, and if M is large this is a large matrix.

This was a mistake, we meant that the number of observation is smaller than the number of unknowns.

Page 3764, line 21: How are the "8 surrounding grid cells" defined in a variable resolution grid?

This is done before the conversion to the variable grid. We have now clarified this in the text.

Page 3765, line 2: Why is the dimension of B PxP? P was introduced on page 3759 as the dimension of the global model grid.

This should be K x K. We have now corrected this.