Dear Reviewer,

Thank you for taking the time to review our work.

In the review it is stated that the manuscript needs changes in order to make the paper easier to follow, to highlight the key points, and to clarify the methodology. We have addressed the suggested revisions throughout the body of the article.

In order to make the paper easier to follow we have rewritten sections (such as section 2), reorganized results and added comments. Some points that were not completely transparent were clarified following also suggestions from the other referee.

The highlighted key points are now underlined in the abstract. This is:

- that our estimated fluxes are broadly consistent with other studies and suggest a possible underestimation of Tokyo Bay Area fluxes in the CDIAC inventory;
- that we examine the impact on our results of using
  - different wind data to drive the model,
  - only sparse surface CO2 data vs. including aircraft observations,
  - and different a priori source fluxes,
- that all of these factors are shown to cause significant differences in the estimated flux, and highlight the challenges in estimating regional CO2 fluxes.

In order to clarify the methodology, the section 2 was rewritten.

Replies to specific comments are given below.

Pg 10624, lines 9-10

The specific values that were quoted in the abstract correspond to a particular mean and variance of the posterior fluxes in an early version of the manuscript – presented in at the AGU fall meeting in 2009. We have made explicit the values within the body of the article. Because it required a more detailed explanation on the dependency on the prior values and other uncertainties, we have moved the values to the body of the article (paragraph 3, section 4.2).

Pg 10625, lines 6-10

The micro-scale flux measurement approach is not feasible for the larger area because of a cost consideration. Given that a typical micrometeorological estimation covers an area of order of 1 km^2, and considering that the area of Tokyo city proper (without including neighboring prefectures belonging to the megalopolis) is 2820 km^2, the cost of a comprehensive study is prohibitive as well as impractical because of the difficulty of installing towers everywhere. It may become possible if in the future the flux measurement become an operational government policy, but this is beyond the scope of the basic research of the current paper.

Pg 10625, line 27

"The Tokyo Bay Area. . ." has been made a separate paragraph. The very high concentration of anthropogenic CO2 makes Tokyo a good benchmark for case studies. The anthropogenic signal is significantly higher than the background and biogenic sources.

Pg 10626, lines 8-11

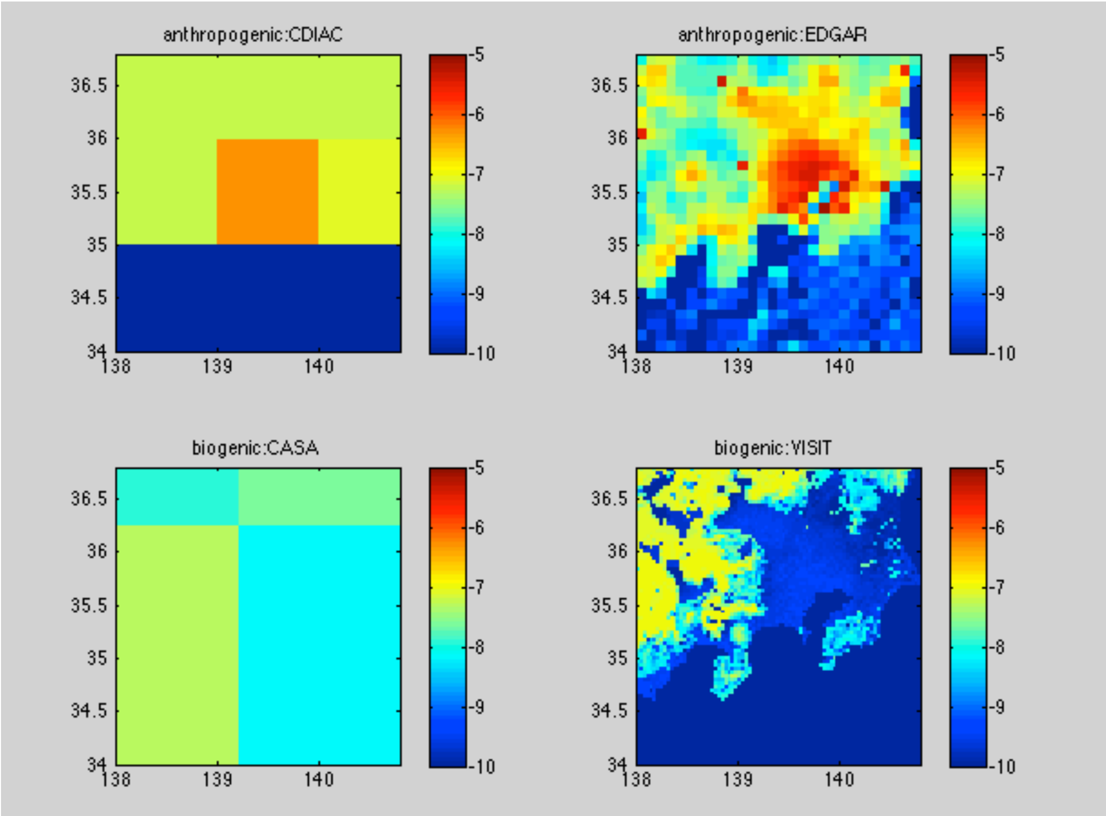
The inventories used where:

- EDGAR (http://edgar.jrc.ec.europa.eu/overview.php?v=42),
- CDIAC, (http://cdiac.ornl.gov/mission.html),
- VISIT (Vegetation Integrative SImulator for Trace gases (Ito, A., Ichii, K., and Kato, T.: Spatial and temporal patterns of soil respiration over the Japanese Archipelago: A model intercomparison study, Ecol. Res., 25, 1033–1044, doi:10.1007/s11284- 010-0729-8, 2010 ; Saito, M., Ito, A., and Maksyutov, S.: Optimization of a prognostic biosphere model in atmospheric CO2 variability and terrestrial biomass, Geosci. Model Dev. Discuss., 6, 4243–4280, doi:10.5194/gmdd-6-4243-2013, 2013.)

and

- CASA (Carnegie AMES Standford Apprach - http://geo.arc.nasa.gov/sge/casa/bearth.html, Potter, C., S. Li, and C. Hiatt, 2012, Declining vegetation growth rates in the eastern United States from 2000 to 2010, *Natural Resources*, doi:10.4236/nr.2012.)

The selected biogenic inventories where consistently over an order of magnitude smaller than the anthropogenic inventories. The figure shows the spatial extent of fluxes in different inventories.



A selected spatial average illustrates our point:

CASA mean	=	-2.3011e-08
VISIT mean	=	2.3950e-09
CDIAC mean	=	2.0579e-07
EDGAR mean	=	7.6056e-07

In this case CASA and VISIT correspond to different winter months for comparison. The spatial averages are taken over the urbanized terrains in the Tokyo megalopolis within the whole selected modeling area. The average biogenic emissions for this period and the location of the center of Tokyo (specifically 35.6848 N , 139.7525 E, the coordinates of the Imperial Palace) are:

- CASA (imperial palace) = -2.0825e-08,
- VISIT (imperial palace) = -3.3313e-10,
  versus anthropogenic emissions of
- CDIAC (imperial palace) = 5.1489e-07,
- EDGAR (imperial palace) = 3.7462e-06,
  a ratio of up to 10 -4.

We believe that under these circumstances other sources of uncertainty prevail, such as the error arising from the PBL height parameterization.

We have referred to inventories referenced in the literature. The question of their reliability is beyond the scope of the present work.

It is certain that the choice if the inventory influences the posterior estimate. This is a fundamental characteristic of Bayesian methodologies. We have shown that the change in the posterior versus prior is larger in CDIAC that in EDGAR, which confirms the hypothesis that CDIAC underestimates the fluxes in central Tokyo (due to its coarseness). The fact that increasing EDGAR yields lower posteriors reinforces the presumption that EDGAR is more accurate than CDIAC for this case.

The references where added to the introduction and the rest of this entry to the discussion.

Pg. 10626, section 2

The section has been modified. We thank the reviewer because the comments have been helpful for a complementary paper describing the method in detail to be submitted to GMD. In particular we have revised to clarify what was used in this study, versus a general discussion of how this might be done.

Pg. 10627, section 2.1.1

The Lagrangian model time step is 15 minutes. The turbulent perturbation is $1/50^{th}$ of this time. We have kept trajectory output every hour. The time resolution of the advecting winds is 1 hour for the regional and 30 hours for the global meteorological models respectively.
The footprints are averaged daily to make the data more manageable. For presentation the data has been also averaged monthly for presentation purposes.

The resolution of the initialization of the trajectories is arbitrary only limited by numerical representation in the computer. There are errors of representation in the space dimension and possible biases associated in the atmospheric model that are beyond the scope of this work, but the time resolution is of data initialization is 1 second. Therefore the measurements are not averaged to one daily measurement and no biases nor uncertainties are introduced in this step. To our knowledge this is the first time such high resolution is applied to the estimate of CO_2 fluxes in this region using CONTRAIL data. The original study dates back to 2009.

Pg 10627, section 2.1.2

The final aim of the whole trajectory calculation is to model the function G(x,t;y,s) (transition probability/Green's function, solution of the Fokker Planck-Kolmogorov inverse equation etc.). There are 2 time dimensions involved: the forward time dimension t and the backward time dimension s. The forward time dimension t is associated with the measurements and its time resolutions is one second for the initialization of the trajectory ensembles (linear interpolation in time, 1 second is necessary for in situ aircraft measurements that are collected continuously, the tower data is already averaged hourly from the data supplier). The backward time dimension is associated with the trajectories whose advection time step resolution is 15 minutes in the Lagrangian model calculation. The trajectory positions are recorded every hour during 3 to 7 days backwards in time from the time of the measurement (in practice the initial positions are ingested by a given model run model in batches every hour, and we have set the system to effectuate one model run every day so every model run ingests 24 batches of initial positions-measurements with a time resolution of 1 second).

Once the position files are available, we proceed to construct the SRR matrix averaging in time and space the trajectories stored in snapshots of 1 hour and with an arbitrary space resolution again only limited by the numerical resolution of the computer number representation. The SRR construction involves particle counting within a space region and a period of time whose resolutions are different from the resolutions involved in the trajectory calculation, We have tested space regions with 0.1 or 0.2 degree resolution and also aggregated following the prefectural administrative limits. We have tested time resolutions for the SRR matrix between 1 and 12 hours with a default time step of 3 hours.

Pg 10627, line 20

The EDGAR a priori fluxes have a 0.1 degree resolution. CDIAC is interpolated to this resolution for purposes of comparison only. In other words: we expect that the results provided by EDGAR are more accurate but it is interesting to compare with a lower resolution model. It is also interesting to test the methodology as a benchmark for future work.

Pg 10627, lines 21-22

Adaptive aggregation means in this context that close to the center of Tokyo the resolution is high, and far away from the center of Tokyo the resolution is low. TBA is in practice the densely populated areas within the Kanto area. We have used GIS to describe the territory as well as the a priori grid high flux grid cells to define the near field. The far field is divided in "rest of Japan", sea grid cells and areas outside WRF domain.

Pg 10627, lines 23-25

"Source geometry specifications . . ." means the specific partition of the geography of Japan for the purpose of constructing a given SRR matrix.

Pg 10628, line 1

We have referenced and described the EDGAR and CDIAC fluxes in the methods section, together with the biogenic models CASA and VISIT.

Pg 10628, lines 5-10

The sensitivity studies are described in the text. The comparison of open ocean vs. model initialization is also included.

In this case, the open ocean site is not a worse background than the global model because the model is scaled with the basis state of the northern hemisphere at this time. The most basic state is an open ocean site. The perturbations from a city plume from continental Asia are likely already mixed with the background. Also, continental perturbations represent only a fraction of the Tokyo signal, which is among the highest in the world. We would use reliable measurements in China if we had access to them. Although we understand that this is only an approximation and we mention that we are intending to perform more sophisticated studies in the future, we argue that other sources of uncertainty impact more in the final results (i.e. PBL height or a priori error). Note that this is the first study submitted using this method and the multiple complexities of the challenging problem force some unavoidable approximations.

See also reply to Pg 10632, lines 1-4 below.

Pg 10628, lines 9-14

The first sentence refers to the method presented here, and hence does not include a reference (the background provided by AGCM is linked to the answer to the question above and Pg 10632, lines 1-4 below). We have expanded on the global model work. The second sentence refers to another work that supports the claim with concrete examples and hence does include a reference (Pisso and Legras, 2008).

Pg 10628, section 2.1.4

The whole section was entirely rewritten and enriched with examples. We hope it is clearer now (this section is the basis for a separate paper describing the method in detail to be submitted to GMD).

Pg 10629, lines 8-11

This term is inspired in Morse and Feshbach (1953) but adapted especially to this case and in that respect can be viewed as original. It is a comment of an approximation always present in geophysical studies but seldom mentioned. Fluid dynamics studies modeling experiments with much more controlled conditions may represent the actual theoretical boundary source term. The resolution of the gradient at the ground is not enough in meso-scale models, in part due to the uncertainties in the representation of the boundary layer.

Pg 10629, lines 12-13

G is the discretized Green's function for the transport-diffusion equation; t and s are the forward and backward time dimensions associated with the Green's function for the transport-diffusion equation.

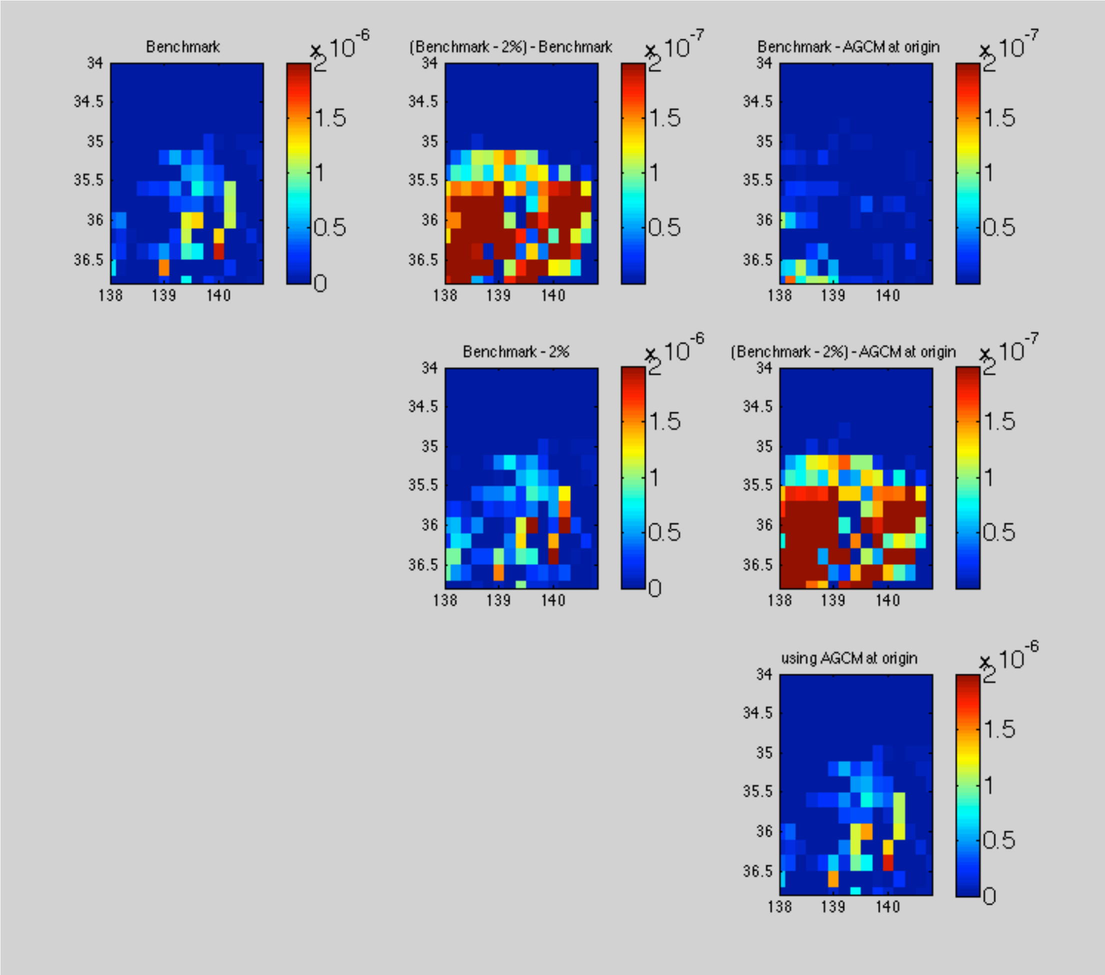
Pg 10632, lines 1-4. (also reply to Pg 10628, lines 5-10)

Mainly because of simplicity and availability of data. We have included the tests performed with AGCM data. This was intended as a separate paper at the time this work was performed (2010/2011). We have developed a sophisticated adaptation of a previously developed method to calculate the background with very high resolution and accuracy (Pisso 2006). Is has been shown elsewhere (Legras 2005) that whenever a quality of the input initial 3D fields is good enough, the reconstructions are accurate and stable. This applies to the present case taking as 3D composition fields the AGCM simulations provided by P. Patra. It has to be taken into account however, that all Eulerian simulation modeling requires a background as well. Frequently the outputs need to be rescaled to fit observations for this reason. To the best of our knowledge the factor that has the largest impact is the choice of the scaling factor. Therefore we chose for many of out tests as an objective background data from a station in the middle of the Pacific Ocean the well-established Mauna Loa time series.

In the Figure, the plots in the diagonal show the inverted fluxes for a day in Jan 2007: (1,1) the benchmark Mauna Loa background; (2,2) the benchmark Mauna Loa background – 2%; (3,3) an inversion performed with the backward Lagrangian diffusive ensembles (BLDE; Legras et al. (2005), Pisso (2006)) connecting with AGCM data (necessarily rescaled to fit the background observations).

It can be observed that the difference between the benchmark Mauna Loa background and the benchmark – 2% (2,1); is relatively large and similar to the difference between the inversion with BLDE and AGCM output with the benchmark – 2% (2,3). In contrast, the difference between (1,1) and (3,3) is small, suggesting that the 2% modification in the benchmark background data prduces a much larger effect than the gain in accuracy provided by the more sophisticated BLDE method.

We argue therefore that in this particular case, the use of an ocean clean air site can be tolerated in terms of error and offers the advantage of simplicity. For this reason we suggested that it was preferable in the ACPD manuscript. We have clarified the choice in the discussion of the final version of the paper.



Pg. 10632, section 3.1

The calibration scale is that of NIES, so large errors due to offsets are unlikely. The calibration error has an effect on the measurement error covariance has been described together with the a priori error covariance. We believe that it is unlikely that the difference off adding CONTRAIL information is due only to the calibration scale. We believe that this has more to do with information absent from the tower dataset. We argue that CONTRAIL data allows to better assessing downwind transport of the city plume.

Pg. 10632, lines 13-19

The Tsukuba tower data is provided hourly aggregated. The data provider (NIES) performs the aggregation (averaging) at the source.

Pg. 10632, lines 20-25

The time resolution of the CONTRAIL data is 1 second and it is not aggregated in any way.

Pg. 10633. Section 3.2

The Carbon Dioxide Information Analysis Center (CDIAC - http://cdiac.ornl.gov/) is the primary climate-change data and information analysis center of the U.S. Department of Energy (DOE). CDIAC is located at DOE's Oak Ridge National Laboratory (ORNL) and includes the World Data Center for Atmospheric Trace Gases.

The Emissions Database for Global Atmospheric Research (EDGAR - http://edgar.jrc.ec.europa.eu) provides global past and present day anthropogenic emissions of greenhouse gases and air pollutants by country and on spatial grid. The current development of EDGAR is a joint project of the European Commission JRC Joint Research Centre and the Netherlands Environmental Assessment Agency (PBL).

The question of the overlap and truly independence of the data inventories is beyond the scope of this work. However they are originated from very different agencies (one European and one American). A caveat has been added.

Biogenic fluxes are small relative to the anthropogenic fluxes comparing averaged values of biogenic fluxes such as VISIT with a priori and a posteriori values of anthropogenic fluxes. Numbers comparing Anthropogenic and biogenic inventories have been provided here and in the discussion.

The convenience of using clean air background sources has been discussed before and in the text also.

Pg. 10634, line 2

T255 is the spectral resolution of ECMWF models (about 80 km in the horizontal direction)

Pg. 10634, line 5.

Months (broadly) covering the whole period of northern hemisphere winter are December, January, February and March. They have been chosen following biogenic processes and PBL activity during this period.

Pg. 10634, line 20

We have described the Kanto area in the introduction.

Pg. 10635, lines 10 – 15

Although both meteorological winds are provided by different centers (NCEP and ECMWF), the data on which these are based (e.g. satellite radiances for the assimilation processes) are not completely independent. Hence, there could be biases in the general weather patterns due to the erroneous model representation of weather systems, fronts and other large-scale atmospheric transport structures. On a smaller scale there could be biases introduced by the limited grid cell size. Sub grid phenomena not represented in the model such as local convection could have a negative impact in the quality of the representation of transport. On the same scale there is a major issue, which is the representation of the PBL. In particular of its height: the night-day difference introduces large uncertainties for flux calculations. Last but not least, the representation of mixing both small scale turbulence in the free troposphere but more importantly the mixing within the PBL can cause biases.

We recognize that all these errors have to be assessed. However we underline that the development of the Eulerian models is beyond the scope of this particular work focused on Lagrangian inversion techniques.

Pg. 10635, line 22

There is data for 2008 and is used in Figure 5. This figure just contains different snapshots of the time evolution to shows the weak trend and to compare with the a priori. We believe that it would be distractive to show an excessive number of figures.

Pg. 10635, line 27

Third party information refers to all the information from different sectors compiled to estimate the inventories (transport, energy consumption, industrial production, economic activity, energy spending etc.)

Pg. 10635, Line 29

We have added the location of the measurement sites to the former figure 3 (now 5).

Pg. 10636, Line 15

The largest power plants (> 1M tons CO2 / year) have been included in the maps in Fig 3. Data from CARMA (www.carma.org)

Pg. 10637, line 4

We agree that in Bayesian inversion systems the choice of a priori flux seems to dominate the a posteriori result, with the observations only slightly nudging the result. However, if the a priori is artificially low, the nudge is towards the higher values, and if the a priori is artificially high, the nudging is towards the lower values. If the a priori and a posteriori fluxes are of the same value ("equilibrium point") and this is not due to the lack of data (which is not the present case) then the inversion supports the hypothesis that the a priori describes the actual flux. However this of course depends on other possible errors and biases in the model and on the interpretation of "the truth". From a purist's point of view the correct statement would be that the observations do not reject the a priori.

On the other hand, this study is not an operational assessment but rather a benchmark of the method. The choice of Tokyo is related to the fact that a number of trustworthy information is available. In particular, it is known that this version of CDIAC emissions underestimated the point-wise fluxes of the city because, as it is intended for global models, the emissions are spread over large areas, overestimating emissions of rural areas to compensate. This is fit to the purpose of the database for initialize global models, but we found worth to mention that such information can be deduced from the method.

Pg 10637, lines 16-22

We are referring to the a posteriori (retrieved) fluxes only. We have reworded for clarity.

Pg 10637, line 24

The a posteriori flux values are displayed in Figure 5.

Pg 10638, lines 1-2

Our retrieved value for the suburban areas of the Tokyo megalopolis was compared to Moriwaki and Kanda's (1 mg CO2/m2/s in winter). The order of magnitude is consistent for areas in the outskirts of the city center – consistent with Kugahara, the area where the study was performed. The values are higher than M&K with EDGAR and lower than M&K with CDIAC. The overall urban region flux would not represent the same kind of emissions since they focused on a small patch of a particular suburban area only and we are including the whole megalopolis. That's the main difference between both studies. We have used emission data collected and processed by EDGAR from 2005 onwards. The study of Moriwaki and Kanda is from 2001. It is possible that the situation changed in the particular neighborhood towards the end of the period of out study (2009).

Pg. 10638, lines 4-5

We included this as a guideline only (order of magnitude of semi urban emissions, as the sum of Tokyo city + surroundings is indeed a mix ox urban, suburban). The park area in Germany is within mixed area. We have suppressed this reference since it may be confusing to the reader as pointed out by the reviewer.

Pg. 10638, lines 10-24

We have included a more detailed discussion of the uncertainties in the inversion process including a posteriori uncertainties. (see also reply to reviewer 2)

Pg. 10638, line 26

The Courant–Friedrichs–Lewy condition (CFL condition) is a necessary condition for convergence while solving certain partial differential equations . It arises when explicit time-marching schemes are used for the numerical solution. As a consequence, the time step must be less than a certain time in many explicit time-marching computer simulations, otherwise the simulation will produce incorrect results. (From Wikipedia)

In this case it means that the time step times the speed of the wind cannot exceed the grid size, otherwise the particles would jump over grid cells with the consequent numerical noise. It is just a numerical accuracy consideration.

Pg. 10638, lines 25-30

The time step of the output is 1 hour. This corresponds to the time step along the ‘s’ dimension or ‘backwards’ as explained before.

The single day run refers to a technical way of organizing the input points in the trajectory model and applies to the time dimension of the data t (forward). The Lagrangian model ingests all data corresponding to one day as initial points, with a time resolution of 1 second. These are in turn taken in batches every hour (keeping the resolution of 1 second). For each of these points, ensembles of trajectories (ranging between 100 and 2000 members depending on the sampling resolution required) are initialized on every individual measurement point. Then these individual particles or trajectories are advected backwards with a time step of 15 minutes. The advectig field is taken form the meteorological (3 hr resolution) or mesoscale (1 hour resolution) and linearly interpolated in time.

Pg 10639, lines 1-10

The robustness of the conclusion is discussed together with the a posteriori flux uncertainty. It is, of course, affected by errors such as the PBL parameterization.

The inversion method can be viewed as a tool for discriminating between inventories based on the information provided y in situ measurements.

To some extent and as it was mentioned before in this reply, it was expected that EDGAR was more accurate because its higher resolution. Again, this paper is intended as a benchmark rather than as an operational system. A caveat has been added.

Pg. 10639, lines 21-27

Although the Keller study assessed a different gas in a different region the underlying methodology is strongly linked (they used also the model FLEXPART) so far only a few studies applying these kind of techniques exist. Also the aim of the technique is to verify reported inventories. The underlying idea is that in the future we will be able to apply the same technique to many other gases. And pinpoint under-reported anthropogenic emissions. We have suppressed the reference.

Pg. 10640, lines 11-14

In the conclusions, we are not claiming that the posterior result are ‘closer to reality’, but rather that the availability of more data ‘provides a stronger constraint’. We are not implying that it is “better” although it is in general accepted that more measurement data are frequently improve the results. It is difficult to benchmark comparing to reality; again, the Bayesian approach would rather refer to a posteriori probability. Notice the last paragraph: “An improvement in the coverage and density of the network distribution is expected to significantly contribute to the overall quality of the inversions. The method may provide an appropriate tool for selecting a priori flux inventories on the condition that sufficient data are collected in adequate locations. Further reduction in uncertainty is required in order to establish a method suitable to inform policy.“

Figure 2. The figure represents an explicit example of a SRR matrix, an important mathematical objet for this study. The rows of this matrix correspond to measurement points (in situ data) corresponding to a certain period of 24 hours. The columns represent emission sources. Red color at an entry of the matrix means that this emission source (column) has impact on the mixing ratio of the row (the measurement). In T1, T2 and T3, T stands for Tsukuba. 1, 2 and 3 stands for the different heights at the tower (at 25m, 100m and 200m). These lines are the same lengths because the sites are hourly averaged and for a 24-hour period all of them have exactly 24 readings, which translates in 24 rows (the same as the upper rows: for the ground sites of Kisai and Dodaira). For every site, the vertical axis can be interpreted as the t time dimension (forward) and the horizontal axis as the s time dimension (backwards) in the Green’s function.

The lower rows correspond to the CONTRAIL data. These data are not averaged and hence there are more data during the 24-hour period (typically between 100 to 200 CO_2 readings). The CONTRAIL data are affected more by the near filed. The blank rows in the middle correspond to measurements in the free troposphere that are not affected by local sources. The fact that the SRR entry value is zero guarantees that the rows are not going to be used as constrains during the inversion process.

Technical comments:

As the reviewer remarks, the measurement data was reported as mole fraction or mixing ratio, not as concentration. We have replaced in the appropriate places.

Pg 10626, line 6. The sentence was divided into 2 shorter sentences.

Pg 10628, lines 22-23. We have modified the sentence.

Pg 10629, equation. The terms x, t, y and s are the position x and time t of the receptor and the position y and time s of the source.

Pg 10638, Line 4. The unit has been added.