Author comments in reply to the anonymous referee on "A tropospheric chemistry reanalysis for the years 2005–2012 based on an assimilation of OMI, MLS, TES and MOPITT satellite data" by K. Miyazaki et al.

We want to thank the referee for the helpful comments, and for his/her compliments on the work presented, the evaluation of the results and clarity of the text. At the same time we fully agree with the referee that the major challenge lies in "unraveling the reasons behind the changes observed" because many aspects of the model are optimised simultaneously in the assimilation process. We have revised the manuscript according to the comments, and hope that the revised version of the manuscript is now suitable for publication. Below are the referee comments in italics with our replies in normal font.

Reply to Referee #1

Overall, the results of the assimilation are quite encouraging. Nevertheless, unravelling the reason behind many of the observed changes proves challenging, given the tremendous number of degrees of freedom for an inversion such as this and the sheer volume of results to analyze. I wish there were more instances where the assimilation results could be translated directly into an improved understanding of some underlying deficiency in the model's transport, chemistry, or deposition. Further, there are a few areas where the inversion doesn't preform as well (such as for NO2 concentrations in polluted areas) or where the results just don't make sense (such as the inference of large increases in NOx and CO emissions in the US and Europe). The paper will be suitable for publication after the authors make revisions to address these and other comments provided below.

The revised manuscript more explicitly discusses the limitations of the current chemical reanalysis calculation, as described below.

Specific comments:

242: It seems a bit out of place to bring up advantages of this approach over another here. But if you want to discuss this, then it also seems a bit odd to only mention the advantages of one approach – surely there are disadvantages as well.

The sentence has been removed from the manuscript.

260: It might be clearer to say: Xb is the ith row (or column? I'm not sure) of an N × k matrix Xb,

where . . .

Corrected.

260: What is the size of N here? Is it just the size of the state vector (38), of the state vector times the physical system dimension?

The size of N is the state vector times the forecast model dimension, which is described in the revised manuscript as follow:
'... where N indicates the system dimension (the state vector size times the physical system dimension)'

Eqn 3: To be consistent with Eq (1), shouldn't it be Yib?

Corrected.

Section 2.2.1: Something is missing from this section, namely the application of the forecast model itself. Where does that come into play? Should it act on xb, or xa from the previous step?

The following sentence has been added to Section 2.2:
'The assimilation step transforms a background ensemble ($¥vec{x}_{i}^b; i=1,¥ldots, k$) into an analysis ensemble ($vec{x}_{i}^a; i=1,¥ldots, k$) and updates the analysis mean, where x represents the model variable, b the background state, a the analysis state, and k the ensemble size.'

The final sentence in Section 2.2.2 has been rewritten as follows:
'The new ensemble members x_i^b after the next forecast step are obtained from model simulations starting from the analysis ensemble x_i^a.'

Section 2.3: The explanation of the state vector is not clear. Line 300 implies that the following discussion pertains to both emission and concentration scaling, but then the description that follows on lines 302 - 304 is only for emissions scaling.

The sentence has been rewritten as:
'The chemical concentrations in the state vector are expressed in the form of volume mixing ratio, while the emissions are represented by scaling factors for each surface grid cell for the total NOx and CO emissions at the surface (not for individual sectors), and for each production rate profile of the LNOx

sources.'

Section 2.5: Why is the notation here for x and y different than other sections, where they are not italicized?

Corrected.

Section 2.7: I realize that inclusion of scaling factors in each surface grid cell for emissions and each grid cell for concentrations is made feasible through the localization step (otherwise the system dimension would be too big). Still, the details of how this is setup are not clear. Are different sets of ensembles used within each localization region, or are there only 30 ensemble members spanning the entire globe? If the latter, this seems like a tremendously large space to span by so few members. The scale L seems to be of order of a few grid cells in the horizontal. But CO emissions have an impact on concentrations several km away. How is the setting of L to only 600 km justified?

The observational information is localized in both the horizontal and vertical to avoid spurious long-range error correlations caused by the limited ensemble size. The localization scale (L) was optimized based on sensitivity experiments on the basis of comparisons with independent observations. L = 600 km corresponds to the cut-off radius of 2191 km (i.e. the localized area diameter of 4382 km), which enables us to assimilate remote observations. The ensemble spread is estimated at every model grid point based on the ensemble model forecast. The background error covariance varies with time and space, reflecting the dominant atmospheric processes and locations of the observations. The sentences have been rewritten as follows:

'The horizontal localization scale L was set to 450 km for NOx emissions and to 600 km for CO emissions, LNOx, and for the concentrations. The physical vertical localization length was set to ln(p1/p2) [hPa] = 0.2. These choices are based on sensitivity experiments (Miyazaki et al., 2012b), for which the influence of an observation was set to zero when the horizontal distance between the observation and analysis point was larger than $2L¥times¥sqrt{10/3}$ (e.g., the cut off radius is set to 2191 km for L = 600 km). We also account for the influence of the averaging kernels of the instruments, which captures the vertical sensitivity profiles of the retrievals. The ensemble members and ensemble spread (error covariance) do vary from one location to the next, and from one species to the next, thereby representing the large number of degrees of freedom contained in the model and they way these are constrained by the observations.'

593: Or this could indicate mis-specification of R?

Yes, there is a possibility that the observation error is not reasonably specified. The sentence has been rewritten as follows:

'For the OMI NO_2 assimilation, the $\chi 2$ is > 1, which indicates overconfidence in the model or underestimation of the super-observation error (computed as a combination of the measurement error and the representativeness error).'

604: This is a bit at odds with the figure and following sentence, which show that χ2 is not constant. Perhaps rephrase?

The sentences have been rewritten as follows:

'Before 2010, the annual mean $\chi 2$ is roughly constant, which confirms the good stability of the performance. Seasonal and interannual variations, especially after 2010, of $\chi 2$ can be attributed to variations in the coverage and quality of satellite retrievals as well as changes in atmospheric conditions (e.g., chemical lifetime and dominant transport type).'

701: One need not hypothesize about the information content of the TES data – the DOFs (trace of averaging kernel) will quantify this directly and could be used to check your explanation here.

The following sentence has been added:

'Jourdain et al. (2007) showed that the TES retrievals have 1-2 DOFs in the troposphere, with the largest DOFs for clear-sky scenes occurring at low latitudes where TES can distinguish between lower and upper tropospheric O_3.'

Section 4.3: One aspect that warrants discussion is the difference between the distribution of the analysis increment and the changes in OmF between the control run and reanalysis. Granted transport, chemistry, and the observation operator make these not correspond 1:1; however, I had a hard time rationalizing what I saw. For example, presumably much of the improvement in MLS ozone occurs in the upper troposphere. How is it that the OmF for MLS, which was mostly negative south of 50 S but positive and negative north of 50 N be improved by an increment that is positive nearly everywhere?

The OmF for MLS O_3 is shown for the vertical layer between 216 and 100 hPa in Fig. 2, whereas the analysis increment is shown for the vertical level at 200 hPa in Fig. 3. Because the ozone concentration increases rapidly with height at the altitudes, the OmF averaged between 216 hPa and 100 hPa is largely

different from that at 200 hPa, and does not show good correspondence to the analysis increment at 200 hPa. The analysis increment averaged between 200 and 100 hPa (as shown below) corresponds well to the OmF averaged between 216 and 100 hPa (Fig. 2 in the manuscript). To explain this, the following sentence has been added:

'The obtained analysis increments correspond well to the OmF in the control run at the same altitude (figure not shown), confirming that the data assimilation effectively reduced the model errors through the analysis steps.'

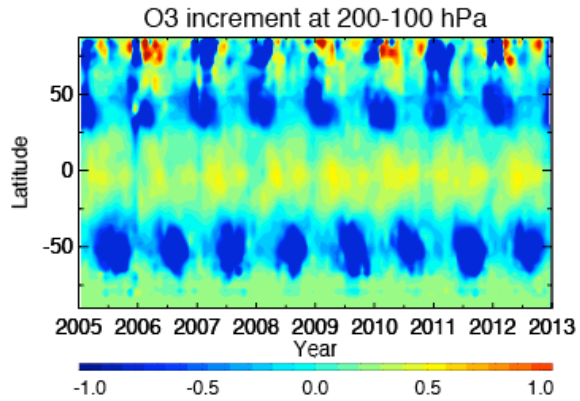

Figure: Time-latitude cross-section of the analysis increment obtained for O_3 between 200 and 100 hPa (in ppb/analysis step).

Also, line 717 implies that the drastic differences in ozone increments at different vertical levels is a sign that the system is working well. I would consider another possibility that the system is under-constrained, despite the results of the χ2 test, and that the increments are exhibiting high frequency oscillations that lie in null space of your forecast model (rather than the observation space). The χ2 test makes assumptions about the normality of the state parameters and observations, which may not hold.

There might be such a possibility. However, in all cases studied, we did not find any high-frequency oscillations in the obtained analysis increment. In data assimilation, those aspects of the system which are not constrained by the observations will stay close to the model forecast if the assimilation is working properly, and we should not expect spurious oscillations. This is similar to retrievals, where a-priori information will suppress null-space oscillations. If we observe clear analysis increments this usually is a good sign and means the assimilation is using the information from the observations to improve the OmA as compared to the OmF. Our results suggest that there is enough information to distinguish emission signatures from concentration biases.

Fig 7: Maybe something is mis-labeled, but I had a hard time following the description of this figure in the text, in comparison to looking at the figure itself, which seems to show pretty much no improvement

between the control and reanalysis for O3. This is further confused by the Table which is referenced describing the errors in ppb, while the text describes them in %, which makes it a bit hard to follow.

The first sentence of the second paragraph in Section 5.1.2 has been rewritten as follows:
'Although the improvement is not large in the upper troposphere (500-300 hPa, Fig. 7), an improved agreement with the MOZAIC/IAGOS measurements is found in the reanalysis run in the middle troposphere (850-500 hPa) and at the aircraft-cruising altitude (300-200 hPa), as summarised in Table~1.'

Because ozone concentration largely varies with height, we believe that it is useful to discuss the relative error (in %) in the main text.

Fig 9: I also found it difficult to identify in these figures the features described in the text. Please instead provide a plot of the model minus observation for the control and reanalysis results.

The model minus observation plots have been added in Figs. 9 and 13.

854: Is the author refering to other previous studies (if so, please cite them) or the present work?

The sentence has been rewritten as follows:
'These characteristics of the bias are commonly found in comparisons with global ozonesonde observations in this study (c.f., Section 5.1.1) and are reduced effectively in the reanalysis.'

879: however, the performance was significantly worse at the surface

The sentence has been rewritten as follows:
'The comparisons show improved agreements in the reanalysis for the middle and upper troposphere during INTEX-B over Mexico and during the ARCTAS campaign over the Arctic, but the model's positive bias near the surface is further increased for the INTEX-B profile.'

931: I'm not sure what is meant by "emissions constraints provided at the ground surface". Also, it's not clear why the improvements with respect to the IAGOS data are attributed to emissions rather than the adjustment of the concentrations directly.

In the reanalysis calculation, the CO concentrations were not adjusted by the data assimilation analysis.

Only the surface emissions were adjusted for CO in the data assimilation framework, as described in the manuscript. The sentence has been rewritten as follows:
'This confirms that the constraints provided for the surface emissions are propagated well into the concentrations of the entire troposphere with a delay in the peak timing and decay in the amplitude. Note that the CO concentrations were not directly adjusted in the data assimilation framework.'

971-977: This is interesting. I wish there were more analysis like this showing how careful analysis of the model performance can be used to parse errors in emissions from errors in concentrations.

We plan to conduct more detailed analyses in future studies to demonstrate the usefulness of the simultaneous concentration and emission optimization.

1052: I wonder to what extent the poorer performance for the NO2 concentrations in urban areas is owing to the adjustments made only to NOx and CO and not VOCs.

I agree that further constraints on wider fields could help to adjust the model chemical equilibrium state and to reduce the negative bias in the NO_2 concentrations. This is discussed in Section 8 as follows:

'For instance, the reanalysis still has large negative biases in NO_2 concentrations over the polluted regions, which may be associated with errors in for instance the model chemical equilibrium states, planetary boundary layer (PBL) mixing, and diurnal variations of chemical processes and emissions. Adjusting additional model parameters such as VOC emissions, deposition, and/or chemical reactions rates by adding observational constraints will help to reduce model errors.'

1055: This argument is a bit dangerous to make, since the perturbations from the model's natural dynamic chemical state are being forced by the state vector increments. So this would seem to be a potential pitfall of adjusting both concentrations and emissions simultaneously.

Indeed we agree that there is a danger of adjusting the concentration even though the emissions may be wrong, or, vice versa, correcting emissions to compensate for chemistry/transport errors. Concentration adjustments may quickly be lost in the PBL due to the short lifetime of NOx, while emission adjustments could be more efficient to store the information over longer time periods. The sentence has been reformulated as follows:
'There may be several reasons for the remaining underestimation of NO_2 concentrations. The analysis increment can partly be lost after the forecast because of the short lifetime of NOx (Miyazaki and Eskes

2013), especially when concentrations are adjusted. Other model processes, such as the diurnal cycle, boundary layer mixing and venting, and the chemical equilibrium at overpass may not be described well. Also, the averaging kernels show a relatively small sensitivity close to the surface, resulting in relatively smaller adjustments in the assimilation.'

1070: The impact of the observations at one hour of the day should be much longer reaching, given the substantial role of NOx on the chemical state. Despite the short lifetime of NO2 itself, this impact can be multiple days, reaching several hundred km.

There could be such longer time influences on NO_2 through the propagation of observational information in the complex chemical system, influencing species like ozone or reactive nitrogen reservoir species. Nevertheless, direct (and local) constraints are considered to be much more important in correcting tropospheric NO_2.

1172: It seems like if this still "remains an important issue" then you might find a more recent paper on the topic than 2003.

The following paper is cited in the revised manuscript:
Stone, D., Whalley, L. K., Heard, D. E.: Tropospheric OH and HO_2 radicals: field measurements and model comparisons.. *Chemical Society Reviews*, 41 (19), 6348-6404, 2012.

Section 6: Several studies have investigated trends in NO2 over the US and Europe in the past decade, with direct inferences for NOx emissions trends being largely negative (e.g., Russell et al., ACP, 2012). Unfortunately, the results don't look anything like the trends shown here with increasing NOx emissions. This is probably the weakest aspect of this paper, along with similar issues for the CO emission trends – how do you reconcile these results with known improvements in combustion efficiency / control technology in developed nations? This needs to be seriously addressed.

We agree with this comment and are very much interested in this topic. We have found that there were similar negative trends in the NO_2 concentrations over the eastern US for 2005-2011 between the OMI observations (-38 %) and the reanalysis (-35 %). The negative trend in the reanalysis is larger than that in the model simulation (-17 %) and is closer to the observational estimates, including the result of Russell et al. (2012) (-32 % for the same period but for a wider area). The estimated surface NOx emissions also showed a larger negative trend over the eastern US for 2005-2012, a larger positive trend over China for 2005-2010, and a larger negative trend over Europe for 2005-2010 in the reanalysis than in the a priori

emissions; the estimated emission trends in this study are similar to those in Tong et al. (2015), Castellanos and Boersma (2012), and Gu et al. (2013), respectively, for the same area/period. Although the mean concentration and the interannual variation of tropospheric NO_2 were generally underestimated in the reanalysis as discussed in the manuscript, the estimated emissions can be expected to provide implications for year-to-year emission variations. Because the main purpose of this paper is to describe the general performance of the reanalysis data, detailed analyses of the year-to-year variations in the estimated emission sources will be discussed in a separate paper (Miyazaki et al., in preparation). This is noted in the manuscript.

References:

· Castellanos, P., and K. F. Boersma (2012), Reductions in nitrogen oxides over Europe driven by environmental policy and economic recession, Sci. Rep., 2, 265, doi:10.1038/srep00265.

· Gu, D., Y. Wang, C. Smeltzer, and Z. Liu (2013), Reduction in NOx emission trends over China: Regional and seasonal variations, Environ. Sci. Technol., 47(22), 12,912–12,919.

· Daniel Q. Tong, Lok Lamsal, Li Pan, Charles Ding, Hyuncheol Kim, Pius Lee, Tianfeng Chai, Kenneth E. Pickering, Ivanka Stajner, Long-term NOx trends over large cities in the United States during the great recession: Comparison of satellite retrievals, ground observations, and emission inventories, Atmospheric Environment, Volume 107, April 2015, Pages 70-84, ISSN 1352-2310, http://dx.doi.org/10.1016/j.atmosenv.2015.01.035.

1335: But there is no improvement visible actually at the surface? Or is the scale of the plot just such that this improvement isn't visible? While I don't doubt that including emissions leads to changes in the mid-trop, it seems it should though at least make an equal or lager improvement directly at the surface.

There are also improvements near the surface. The sentence has been rewritten as:

'At the NH mid-latitudes the changes introduced by optimizing the emission factors improve the agreement with the ozonesonde observation from April to August below about 500 hPa (Fig. 17) associated with the pronounced O_3 production caused by NOx increases; the monthly mean positive bias below about 900 hPa is reduced by 10-15 % in the summer and the negative bias between 900-500 hPa is reduced by 30-50 % in spring and summer.'

The improvement is larger in the middle troposphere than in the lower troposphere. This is probably because of larger spatial variations in the lower tropospheric O3 concentration and the coarse resolution of the model (i.e., larger representative error).

1342 - 1348: Shouldn't they both lead to the same O3 trends, the one which matches the observations, just by different means?

Because the interannual variation in the emissions was poorly represented in the a priori emissions (e.g., the anthropogenic emissions for 2008 were used in the calculation for 2010), the ozone concentration trend between 2005 and 2010 is expected to be unrealistic in the case without emission source inversion. In addition, the decreased number of TES measurements after 2010 makes it difficult to produce the ozone trend, particularly in the case without emission source inversion. Therefore, we can expect different ozone trends between the two calculations.

Table 6: The caption should state what is shown (percent error? ppb?), and if the values are model minus observations or vice versa.

Corrected.

Section 7.4.1: All though I'm not quite sure what the numbers in Table 6 represent (see previous comment), it seems odd that in the NH the differences at 200 hPA (16.3 vs 13.2) would be larger than differences at the surface (0.1 vs - 2.3) for the test using different prior emissions. Why would the impact of changing the prior emissions be greatest in the upper trop? Perhaps the results of this test would be better served by showing a plot of the difference in the posterior emissions between the standard and HTAP-based reanalysis.

The different a priori surface emissions led to differences in the estimated LNOx sources locally (at model grid point scale) by up to 15% at the NH mid-latitudes (the annual global total LNOx source differed by only 2%), which might cause the ozone concentration difference in the upper troposphere especially in the summer. The following sentence has been added:
'The spatial distribution of the estimated LNOx sources is also somewhat influenced by the choice of a priori surface emissions in the NH mid-latitudes (not shown), which led to differences in the agreement with the ozonesonde observation in the upper troposphere at 200 hPa.'

Section 7.6: Why not just report the uncertainty reduction as measured by the posterior / prior error, rather than the ensemble spread?

As described in the manuscript, the ensemble spread is influenced by errors in the model input data, chemical or physical parameters, numerical scheme, as well as errors in the measurements assimilated,

while it also reflects the instability of the tropospheric chemical system. We thus believe that measuring the absolute value of the estimated analysis spread (rather than the posterior/prior error ratio) is meaningful for the evaluation of the analysis uncertainty. The $\chi 2$ test provides further evidence that the ensemble spread is reasonable.

Section 7.7: I found the first half of this section (1620 - 1639) quite speculative. It could be removed, given the paper is already quite long.

The chemical reanalysis is still a challenging matter. However, its usefulness may not be well understood by many readers. We thus believe that this section is important and worth keeping, although it adds to the paper's length.

Conclusions: While the overall results of this work are indeed impressive, I feel there were a few aspects which need to be made more transparent, such as the inability to improve surface NO2 concentration in polluted areas, or the strange trends in NOx and CO emissions in the US and Europe. A bit more balanced evaluation of all strengths and weaknesses would be good.

To describe the weakness of the present study, the following sentence has been added:
'For instance, the reanalysis still has large negative biases in NO_2 concentrations over the polluted regions, which may be associated with errors in for instance the model chemical equilibrium states, planetary boundary layer (PBL) mixing, and diurnal variations of chemical processes and emissions.'

The estimated emissions show negative trends over the US and Europe, which will be described in a separate paper (Miyazaki et al., in preparation). Please also see my reply above.

4: retrieval data → data
53: to develop → in developing
55: information, → information
56: the estimates → estimates
121:),by→)by
247: the analysis performed → performing the analysis •
throughout: number % → number%
578: no new paragraph
619: too large → excessive
1151: while → while it

Corrected.