Response to the first reviewer

We thank the referee for his/her positive review and for the provision of useful comments and suggestions. Below we answer them to our best ability. The reviewer comments are in italic. Our responses are in regular font, and changes to the manuscript are given in bold.

The manuscript 'An inter-comparison of tropospheric ozone reanalysis products from CAMS, CAMS-Interim, TCR-1 and TCR-2' presents a description and extensive evaluation of tropospheric ozone from four recent global chemical reanalyses: CAMS-iRean, CAMS-Rean, TCR1 and TCR2. The study performs very detailed comparisons between the reanalyses and independent observations of surface, profile and column ozone and assesses the relative performance of the reanalyses. This includes some very nice analyses of specific aspects of the reanalyses such as the representation of the diurnal ozone cycle. I am really impressed by the amount of work that went into this study and I applaud the authors for their very thorough and well-organized analysis. This type of paper is not easy to write but it is very important for the scientific community, especially as the understanding of the importance of chemical reanalyses is growing.
The paper is well thought out, well-written and well-organized. It is certainly worth of prompt publication. I have only several minor suggestions for edits and some technical corrections.

Minor general comments

1. Tables and the discussion of reanalyses' performance: it would be really good to have the RMSE values shown in percent in addition to absolute values (ppbv, DU). Having absolute values alone makes it difficult to judge if the RMSE is large or not. I appreciate that sometimes, particularly when the mean ozone is low, large percent values may be misleading but that shouldn't be an issue if both, absolute and relative RMSE is shown. In my specific comments I point to some places where having percentages would be particularly useful but it would really be best to have them in all the tables.

We now add barplots with normalized mean bias and normalized standard deviation (instead of RMSE, as per a comment from the second reviewer) in the supplementary material.

2. It would be helpful to have a schematic figure, similar to Davis et al. 2017 Fig. 1, showing the ozone observations assimilated by each reanalysis, indicating whether a bias correction was applied or not, and, as an added benefit, showing time periods of the reanalyses.

We have checked the Fig 1. in Davis et al. (2017), but find it unpractical to introduce a similar figure for our purposes, as this implies considerable overlap with Tables 2-4. Also not only information on the satellite instrument is important, but also the version specification, which implies that the figure cannot replace the existing tables. We now introduce a separate section in the manuscript to discuss any changes in the observing systems.

3. The authors often use the word 'model' as synonymous with 'reanalysis', e.g. L273, L307, L316, L318, and in many other places. I suggest limiting the use of this term to the instances where you are really talking about a model (e.g. 'model levels' or *‘chemistry model’, etc.)*

The reviewer is fully correct. We have checked the manuscript, and replaced ‘model’ with ‘analysis’ or similar, where appropriate.

4. Section 5 contains detailed discussions of reanalyses comparisons against multiple data sets. It’s easy to lose the big picture in all these details. It would be really helpful to include 2-3 sentence summary highlighting the key results at the end of each subsection as it is already done in Subsection 5.3.

We now introduce a summary section at the end of Sec. 4.3, concluding the evaluation against ozone sondes. Likewise we now introduce summary statements at the end of Sec 5.2 an Sec 6:

End of sec. 4.3:
In conclusion, evaluation against ozone sondes has revealed the following:
- The updated reanalyses show on average improved performance compared to the predecessor versions, but with some notable exceptions, such as an increased positive bias over the Antarctic in CAMS-Rean versus CAMS-iRean. Over the Antarctic the TCR-2 strongly improved upon TCR-1, despite the lack of direct observational constraints.
- For individual regions or conditions CAMS Reanalysis and TCR-2 show different performance, but averaged for all regions of similar quality. Best performance, in terms of mean bias, standard deviation and correlation, for the updated reanalyses is obtained for the Western Europe, Eastern US and SH mid latitude regions (both normalized mean bias and standard deviation below 8% at 850 and 650 hPa). Relatively worst performance is found for the Antarctic region, with normalized standard deviation up to 18%. This is likely associated to the fewer observational constraints in the polar regions compared to the other regions.
- In terms of temporal consistency, the CAMS Reanalyses show degraded performance over the polar regions during 2003 and 2004, due to lower quality MIPAS and SCIAMACHY data usage. CAMS-iREAN also shows a change in performance statistics in the polar regions from 2014 onwards, associated to a changes in the MLS retrieval product versions. Furthermore, both CAMS-Rean and CAMS-iRean are affected by the change in the SBUV/2 product versions in 2013.
With the reduced data-availability from TES from 2010 onwards the TCR tropospheric ozone products show changes in their performances. Remarkably, TCR-1 and TCR-2 show overall slight improvements from 2010 onwards. This is marked by reduced positive biases in the lower troposphere over NH-mid-latitude regions and may be attributed to biases in the TES retrieval product, combined with changes in the OMI product, see also Sec. 2.5. Additional Observing System Experiments (OSEs) are needed to identify the relative roles of individual assimilated measurements on the changes in reanalysis bias.
end of sec 5.2:
In summary, CAMS-Rean shows the best ability to capture the regional mean surface ozone and its variability, while particularly TCR-2 (and to lesser extent also TCR-1) shows positive biases and reduced correlations. Particularly good performance is seen over the western US (R=0.95, MB=-0.2), while over east, and particularly southeast, Asia the performance is poorest.

end of sec.6 :
In summary, all reanalyses capture the synoptic to diurnal variability, as illustrated by the assessment of the heatwave event in July 2006. Still there are considerable differences in

performance, depending on the reanalysis, region and season. While CAMS-iRean and CAMS-Rean perform mostly similar, for TCR-2 a considerable improvement was found compared to TCR-1. Overall better temporal correlations are obtained for the summer period compared to winter, and also for Western Europe compared to the Mediterranean region. Further improvements can be obtained by a better description of surface processes, including emissions and deposition, together with higher spatial resolution modelling.

Specific comments and technical corrections

L35 climate-change à climate change

Changed.

L34-36. This sentence conflates two different things: (1) the importance of ozone forcing for climate and (2) a lack of impact of improved ozone representation on long-term weather forecasts. I suggest splitting it into two sentences.

Thanks for this suggestion. We now write:
Owing to its radiative effects, tropospheric ozone is an important driver in climate change (Checa-Garcia et al., 2018). Also it may affect long-range weather forecasts, even if in evaluations no improvement has been detected so far (Cheung et al., 2014).

L38. This deserves more references than just the two that are provided.

This is correct. We now have added references to Monks et al. (2015), Huang et al., (2017), and Hsu and Prather (2009).

L120-121 'to evaluate their fitness for purpose for the various types of application described above'. This sounds a little awkward. Please, consider rephrasing.

We now write:
To assess the quality of these reanalysis products, with attention for the various potential types of application described above, this study evaluates tropospheric ozone…

L158. Was there any kind of bias correction applied to these ozone data, as in CAMS-REAN? Maybe I missed that information. As I stated in my general comments a figure summarizing all these data types and how they're used in each reanalysis would be useful. This information could also be added to tables 2, 3 and 4.

We now explicitly mention the bias correction settings in any of the reanalyses. The settings for CAMS-iRean and CAMS-Rean are identical (variational bias correction for OMI, SCIAMACHY, GOME-2, and anchoring for SBUV/2, MLS and MIPAS), while in TCR-1 and TCR-2 all observations were used without bias correction.

Table 2. For the profile data types it would be helpful to include the vertical ranges or at

least the lowest levels assimilated.

Profile data from MIPAS and MLS instruments in the ranges 0.1 -150 hPa (MIPAS) and 0.1 - 147hPa (MLS) is used. For SBUV and GOME (ERS-2) the vertical resolution is very low, implying that they can effectively be considered as total column retrievals. We now include such a comment in the manuscript.

L200. Why couldn't they be filtered out?

These OMI row anomalies could not be filtered out because at the time this information was not available in the BUFR data which are used as input to the IFS data assimilation system. This had unfortunately not been noticed before running the reanalysis. This information will be taken into account in any future reanalysis. We now write:

-Different behaviour of OMI data between 2009 and 2012, associated to a deterioration in the OMI row anomalies (Schenkeveld et al., 2017) which unfortunately have not be filtered out in the CAMS assimilation procedure;

L231 & L247.Livesey et al., 2011 is not in the reference list. If this is the MLS version 4.2 data quality and description document then its latest version is from 2018 (https://mls.jpl.nasa.gov/data/v4-2_data_quality_document.pdf). Is the MLS data quality screening based on some earlier guidelines? Note that v4.2 didn't exist in 2011.

The reviewer is correct: Version 4.2 was actually used in both TCR-1 and TCR-2. The v3.3 MLS data was used in a predecessor of TCR-1, not assessed in our manuscript. MLS data quality screening is also based on the v4.2 guidelines. We have updated the manuscript, and reference on this.

L245-251. TES should also be mentioned here for completeness.

We now include such a sentence, thank you for this suggestion.

Table 4. According to the table TCR-1 uses MLS v3.3. It's version 4.2 in the text (L230).

The reviewer is correct: it should have been version 4.2, as was already mentioned in the text. This is now also updated in the table.

*L273. Data has been collocated à data **have** been collocated*

Updated, thank you.

L299 'any of the reanalysis model resolutions is considered too coarse' please correct the grammar

changed into:
…because none of the reanalysis model resolutions is considered sufficient to resolve …

L310. What's the frequency of EMEP data?

EMEP provides hourly observations. For our evaluation we use a reference three-hourly time frequency. We now clarify these time frequency aspects specifically in Sec 3.3

L325. I think ‘multiannual’ is one word. At least, please be consistent; ‘multi-annual’ it’s hyphenated a few lines below.

We now consistently write ‘**multiannual**’

Figure 2 appears to be repeated or at least I can’t discern any difference between the top two and bottom two rows. In addition, please, explain in the caption what ‘Season: AYR’ means (‘all year’?) or remove it from the legend.

The reviewer is correct about the duplication, we apologize for this. The reference to ‘Season: AYR’ (referring to full multiannual averaging as compared to seasonal averaging) is removed from the figures.

L325-342. What about the large discrepancy between the sondes and all the reanalysis near the surface at NH subtropics and, to a lesser extent, the tropics?

The near-surface discrepancy for the NH-subtropical region can mostly be attributed to positive biases in any of the reanalyses against the Hong Kong (114.2° E, 22.3° N) sonde observations, see also Figure R1 below. O_3 at the Hilo (155° W, 19.4° N) and Naha (127.7° E, 26.2° N) stations perform much better at these low altitudes. Likewise, for the tropical region a large bias could be attributed to the Kuala Lumpur, Malaysia station (101.3°E, 2.7°N). But due to the sparseness of the observations in these regions it remains difficult to derive general conclusions.

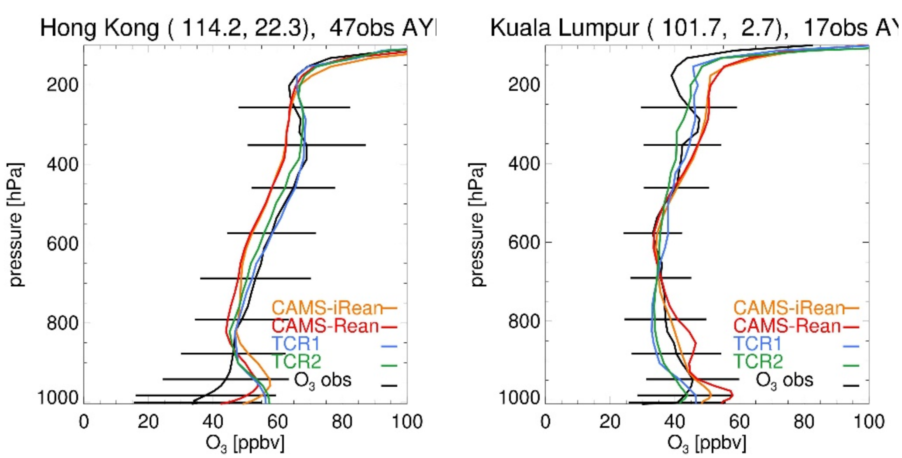

Figure R1. Evaluation of multi-annual mean ozone from all reanalyses sampled at the Hong Kong (left) and Kuala Lumpur (right) stations.

In the manuscript we now write:

In the NH subtropics and the tropics regions the reanalyses show some larger deviation against sonde observations at lower altitudes, which was traced to comparatively large biases at the Hong Kong and Kuala Lumpur stations. Note that in these regions the ozonesonde network is sparse, while the spatial and temporal variability of ozone is large, which limits our understanding of the generalized reanalysis performance (Miyazaki and Bowman, 2017).

L341-346. The comparison with ACCMIP would be easier to see if the biases shown here *were given as percentages in addition to absolute values.*

We now also report on normalized biases (and standard deviation) in new figures in the Supplementary Material. We include a statement on the maximum normalized (absolute) mean bias being below 10%.

L370. Could you briefly justify the use of an 'ozonopause' rather than more commonly used lapse rate or dynamical definitions of the tropopause? In addition, because of the high vertical resolution of ozonesondes they're likely to attain 150-ppbv threshold at very different (and somewhat random) altitudes than the reanalyses. How does that impact these comparisons?

In line with the comment from reviewer #2 we now use a more clear definition of the tropospheric column. We now compute this as the partial column from the surface to 300 hPa. Indeed, this helps to intercompare the reanalyses, as alternatively the altitude of the tropopause level changes between reanalysis.

At the start of Sec 4.2 we now write:
Collocated partial columns from the surface up to 300 hPa, hereafter for brevity referred to as 'tropospheric columns', have been compared to partial columns derived from the sonde observations.
All figures and reporting on error statistics has been updated accordingly.
L373. SH midlatitudes also look messy, especially TCR reanalyses. The absolute RMSE may be less than at high latitudes but relative to the mean column it looks quite high. Here and elsewhere it might be helpful to provide percent values for the mean biases and RMSE.

Following the reviewer's recommendation we now also compute the normalized biases. They indeed indicate difficulties over the SH mid latitudes, although smaller compared to the high latitudes. We now write:

Over the SH mid latitudes the reanalyses show similar features as over the Antarctic, with normalized mean biases within -1DU (-5%, CAMS-iREAN) and 1.5 DU (+10%, TCR-1). The normalized standard deviations over the SH mid latitudes are within 7%, marking a considerably better ability to capture temporal variability than over the Antarctic.

*L390. 'These figures'. It's **one** figure (multiple panels)!*

Changed to '**the panels in this figure'**, thank you.

Figure 4. The caption says that ozonesondes are shown in black but since what's shown is biases w.r.t. the sondes the latter are not really shown at all, are they? I suggest deleting that sentence. Also, please state that numbers of observations are shown as gray dashed lines, even if it's obvious from the previous figure. As a side note, I'm not against multipanel figures but I don't think I've seen one with 21 panels before.

The reviewer is correct: the sentence is deleted now, and explanation of the gray dashed line is included instead.

L401. Why is MIPAS relevant to the troposphere? Is that an indirect impact of assimilating total ozone with stratospheric ozone constrained by MIPAS data? The same question applies to line 451-453 (Antarctic ozone).

The reviewer is correct. To explain this better, in Sec 2.1 the following sentence is included:

Profile observations from limb instruments (MIPAS and MLS) are used to constrain the stratospheric contribution of the total column. In combination with the assimilated total column retrievals this implies that also the tropospheric part is constrained (Inness et al., 2013).

And in sec. 4.3, when discussing the impact of MLS:

Combined with total column retrievals, assimilation of such stratospheric profiles has been shown to also affect the tropospheric contribution (Inness et al., 2013).

L430-433. Any idea what happens around 2010-2011 that causes this improvement over Japan?

It is very difficult to attribute the change in bias statistics over Japan around 2010 for the four reanalyses. Aspects that play a role are following:

- The ozone observations at 650 hPa show relatively large annual mean values, during 2010 and 2012, see Figure S1 in the (original) supplementary material. This may be associated to the increased NOx emissions from China in the preceding decade (e.g. Verstraeten et al., 2015), which show a maximum during 2011 – 2014 (van der A et al., 2017). Note that in the CAMS reanalyses NO_x emissions are not optimized in the data-assimilation procedure, although NO_2 tropospheric columns have been assimilated. Instead an annual trend is assumed in the MACCity based emissions.
- The TCR-based reanalyses show a significant change in their characteristics after 2010 due to a reduction in TES retrievals, which stopped completely after June 2011.
- Both the TCR and CAMS reanalyses are affected by the row anomaly issue in the OMI O_3 (relevant to CAMS-REAN and CAMS-iREAN, particularly during 2009-2012, Inness et al., 2017) and NO_2 retrieval products (relevant to all reanalyses).

It is unfortunately beyond the scope of this work to assess the partial contributions of these effects. To provide more clarity, in the manuscript we now write:

The changes in performance statistics for all reanalyses likely have multiple causes. This includes trends in the observed ozone (Verstraeten et al., 2015), associated to changes in Chinese precursor NO_x emissions (e.g. van der A et al., 2017). Also changes in the observing system are important to consider, particularly the reduction of assimilated TES measurements in TCR from 2010 onwards, and the row anomaly issues affecting assimilated OMI O_3 and NO_2, see also Sec. 2.5.

L445-446. The CAMS reanalyses show some large departures before 2005, especially at 382 hPa. Can you comment on that?

We attribute this to similar causes as identified for the NH polar region, namely the use of early SCIAMACHY and NRT MIPAS O3 retrievals, which are of poorer quality than the OMI MLS observations which have been used from August 2004 onwards, and reprocessed MIPAS data used from January 2005 onwards. We now write accordingly:

Furthermore, CAMS-iRean and CAMS-Rean suffer from relatively large negative biases before 2005, particularly at 382 hPa. This is attributed to similar causes as have been discussed for the Arctic region.

L507. But it's not exactly the same period, is it? Figure 6 shows aggregated data from 2005 to 2012 and S3 is extended through 2014.

That is correct, therefore we provide the exact time range in all the table and figure legends. We now also specify this additionally in this particular sentence.

Figure 8. The caption says 'left' and 'right'. It should be 'top' and 'bottom'. Alternatively, the panels could be labelled.

We now change to 'top'/'bottom', thank you.

L652. 'Figure 11'. I think it should be '12'.

The reviewer is correct, this is now changed, thank you.

L727. Here and elsewhere, please provide percentages in addition to absolute values. How large (small) is 6 ppbv in this case?

We now include such assessment here. We write:
Normalized to local mean O_3 from the CAMS Reanalysis, the standard deviation values at 850 hPa reach 20% over Australia and up to 50% over South America and Central Africa. At 650 hPa these maximum ratios decrease to approx. 10% (Australia) and 20% (South America and Central Africa).

L806. Could you expand on this? It would be very helpful to include a paragraph with specific recommendations for the users: What kind of studies are these reanalyses good for? Which reanalyses are recommended for a particular type of study and which ones are less reliable? Are there any types of problems for which these reanalyses are not useful? This is partially addressed in the second to last paragraph where the authors delineate some issues related to trend and long-term variability studies using reanalyses but I think this type of discussion could be expanded to other areas.

The reviewer is correct that such suggestions could be useful. We now include the following sentences:

The well-characterized, small mean bias in tropospheric columns in these reanalyses suggest that they can be used to provide a climatology of present-day tropospheric ozone. This may serve as a reference for the present-day contribution of tropospheric ozone to the radiation budget, or may provide a climatology for a-priori ozone profiles as required for satellite retrieval products (e.g., Fu et al., 2018). The ability of the CAMS Reanalysis to capture the variability of (near-)surface ozone on multiple time scales, and for many regions over the globe, indicates it is fit for use as boundary conditions for hindcasts of regional air quality models.

L 810. Do you really mean 'any' models or is it 'many' models?

We refer to the model configurations discussed in our evaluation. We now rewrite to:
The relatively coarse horizontal resolution in any of the global reanalysis configurations could also cause significant errors at urban sites.

References

Davis, S. M., Hegglin, M. I., Fujiwara, M., Dragani, R., Harada, Y., Kobayashi, C. et al. (2017). Assessment of upper tropospheric and stratospheric water vapor and ozone in reanalyses as part of S-RIP,. Atmos. Chem. Phys., 17, 12743-12778, https://doi.org/10.5194/acp-17-12743-2017

Verstraeten, W. W., Neu, J. L., Williams, J. E., Bowman, K. W., Worden, J. R., and Boersma, K. F.: Rapid increases in tropospheric ozone production and export from China, Nat. Geosci., 8, 690–695, doi:10.1038/ngeo2493, 2015.

van der A, R. J., Mijling, B., Ding, J., Koukouli, M. E., Liu, F., Li, Q., Mao, H., and Theys, N.: Cleaning up the air: effectiveness of air quality policy for SO2 and NOx emissions in China, Atmos. Chem. Phys., 17, 1775–1789, https://doi.org/10.5194/acp-17-1775-2017, 2017.