Reply to Anonymous Referee #2

We thank the anonymous referee for her/his constructive comments and suggestions, which helped us to improve our manuscript. We have carefully considered each of the comments and have modified the text accordingly. Please find below our reply (blue text) to each comment (black).

I find it quite difficult to follow the differences in the 4 different model simulations. It would be helpful to have a table that lists all the relevant differences. For example, yearly-varying emissions versus non-yearly varying emissions and which inventories used, driving meteorology (for nudged) or stating free-running, climate feedbacks or not, etc. This is done slightly in Table 2, however a more comprehensive summary table would be useful.

The requested information is provided in Table S1 in the Supplement, as mentioned at the beginning of Sect. 3 in the paper.

For temperature (figure 1), can you say why the free running models perform better at high altitudes compared with the nudged? This seems to coincide with the regions where the nudging is not applied (above 97 hPa)? What is the difference between the way the climate variables are calculated between the nudged model outside of the nudging region and the free running model?

There is no difference in the way the climate variables are calculated inside and outside the nudged region. At high altitudes (i.e., at the 30 and 5 hPa levels), all model configurations are essentially free-running, since nudging is only applied up to 97 hPa. Nevertheless, the region above is affected by the nudging below (e.g. through altered wave dynamics) and the different sea surface temperatures. For instance, Jöckel et al. (2006) showed that the exceptional polar vortex split observed over Antarctica in 2002 could be reproduced by the model, although the nudging was only applied up to 200 hPa. In that case, the "nudging" improved the result outside the nudged region, whereas here the free running model performs better with respect to temperature above the nudged region. Thus, it obviously depends on the quantity or the feature you look at, and it cannot be concluded that the nudging always improves the results outside the nudged region. A possible mechanism could be that the nudging perturbs the radiative balance which feeds back on the temperature, but identifying the actual reason would require a detailed analysis of the model dynamics and a suite of additional sensitivity simulations, which are beyond the scope of this study.

Why is the nudging not applied throughout the whole model domain?

The nudging is not applied throughout the whole vertical model domain, because previous EMAC studies (Jöckel et al., ACP, 2006 and Lelieveld et al., ACP, 2007) show that a better representation of the stratosphere can be achieved if the nudging is applied up to ~100 hPa or ~200h Pa and not further above. In addition, the boundary layer is not nudged to avoid potential spurious effects (w.r.t. the stability of the boundary layer and the hydrological cycle) resulting from different boundary layer schemes of EMAC and the IFS model (from which the nudging data are). We refer to these two papers in the revised text for a better explanation of this point.

Is the cold bias in the nudged models (seen at 200 hPa in figure 2) due to differences between ERA-operational and ERA-Interim? I wonder how this compares to other non reanalysis temperature datasets? Maybe there is something in the literature? Dee et al., (2011) - The ERA-Interim reanalysis: configuration and performance of the data assimilation system shows a larger RMSE in ERA-operational than ERA-Interim at 200 hPa when compared to radiosondes.

As mentioned in the reply to Reviewer #1 (and now also in the revised manuscript), the global mean temperature has not been nudged, thus we do not expect a bias reduction between the free and the nudged mode. Nevertheless, the suggested reference shows that the difference between the RMS forecast error produced by ERA-Interim and the ECMWF forecasting system that was operational in 1989 is only about 0.2 K at 200 hPa. Although this value refers to the year 1989, we can argue that the cold bias at 200 hPa is not due to a difference in the two datasets. We have added this to the text (see also the reply to a similar comment from Reviewer #1).

For upper tropospheric ozone, you mention that differences between EVAL and QCTM can be explained by differences in lightning emissions. I noticed that lightning emissions are also similarly high for the free running simulations, however, you don't mention this in the section 6.2.2 discussion.

The free running models also have high ozone in figure 13 and 15 in some regions. Do you think this is related? Also EVAL and QCTM have different aircraft emissions, could this also be having an impact? What is the difference in the amount of emitted aviation NOX for these two model runs?

We added the information about lightning emissions from the free running simulations to the text. The fact that TS2000 and ACCMIP have lightning emissions comparable to EVAL2 and comparably high tropospheric ozone further supports the argument in Sect. 6.2.2.

On the other hand, aviation NOx emissions are quite similar in the four simulations: about 1.4–2.0 Tg(NO)/yr for EVAL2 (transient) and 1.8 Tg(NO)/yr (year 2000) for the others, therefore it cannot explain the differences in tropospheric ozone between EVAL2 and QCTM. This has also been added in the revised manuscript.

For OH, it is still useful to quote the global mean tropospheric OH concentration for the different simulations as these can be used to compare the oxidative capacity of different models. Please calculate and add a table giving the mass-weighted tropospheric OH. It would also be useful to compare to other multi-model mean values quoted in the literature. It would be particularly interesting to know the global mean OH calculated for the ACCMIP-S2 run to know how much this new channel is reducing OH. I assume it is by quite a bit due to the impact seen on CO. You state in your conclusions that the improvement in UT O3 due to this channel supports the need for it to be included in models, knowing the impact on OH could further support or counter this. You can compare your values to global mean OH concentrations constrained by methyl chloroform.

Globally-integrated OH concentrations are hard to interpret, given the different methods used in the literature to define this quantity (see, e.g., Lawrence et al., ACP 2001). Lawrence et al. found that "the airmass-weighted and volume-weighted $[OH]_{GM}$ values are generally poor indicators of the global atmospheric oxidizing efficiency with respect to gases such as CH_4 and CH_3CCl_3 with a strong temperature dependence in their reaction with OH." To address this comment, we have removed this sentence and have added instead a more robust discussion on methane and methylchloroform (CH_3CCl_3) lifetimes in Secs. 6.2.4 and 6.2.5, which are much better indicators of the tropospheric oxidation capacity.

Pg 6551, L16: give example of 'climate variables' (e.g. temperature).

Done.

Pg 6552, L1: 'stratospheric input' - of what? Do you mean being transported into the troposphere from the stratosphere?

We mean "stratospheric input of ozone" and have clarified this in the text.

Pg 6553 and Pg 6555: You describe/state you use different versions of MESSy in several places. You can remove initial mention of this on pg 6553 (L18-19). When I read this section I was wondering why you were using two different versions of MESSy as you don't say why. It later becomes apparent that this is because you are using another model simulation that has been done for a different study. Try and combine these two descriptions somehow to make less confusing.

We added a motivation for using two different versions of MESSy at the beginning of Sec. 2. This should make the description less confusing.

Pg 6554 L8-L9: I find it hard to know what you mean here.

We deleted this sentence, as it is not relevant for the discussion.

Pg 6554 L22-L23: Mention use of offline aerosol after you mention heterogeneous reactions, before description of convection at L16.

Done.

Pg 6554, L20: Can these be used to perturb the climate/dynamics of the model?

Yes. With the exception of the QCTM setup, there is a feedback between radiation and dynamics in the model. This has been clarified in the text as follows: “Therefore these constituents are consistently used for the coupling between chemistry and dynamics in both directions via radiative forcing and tracer transport.”

Pg 6556: Why do you apply scaling to ship emissions to estimate transient changes in emissions but don’t do same to the anthropogenic emissions? I feel it would be best to be consistent in your treatment of emissions. Why do you not use the road traffic emissions from the Lamarque et al. dataset?

The evaluated simulations were performed as part of different projects, which motivated their specific setups (e.g., the QCTM run was part of a study on shipping impacts on tropospheric chemistry, while the ACCMIP runs was conducted as part of a large international effort). We added the following statement at the beginning of Sect. 3 to clarify this: “The four simulations were conducted as part of various projects. The specific requirements of each project (e.g., ACCMIP) motivated the different configurations that were applied.”

Pg 6557: Why are the aviation emissions tuned lower?

The aviation emissions are not tuned lower, but the lightning emissions are. As stated in the previous comment, the different setups of the simulations evaluated in this work were motivated by the projects for which the experiments were conducted.

Pg 6560, L12: How do you choose the ‘reference’ dataset? For example, the one with the lowest measurement uncertainty? Or the one with the biggest spatial/temporal coverage? Or maybe more simply this is what is already implemented into the ESMValTool?

The ESMValTool flexibly allows for different datasets. We followed a similar strategy as Glecker et al. (2014), who selected ERA40 as reference and NCEP as an alternative for most of the diagnostics. However, the conclusions of this study are not affected by the choice of the reference vs. the alternative, as the results of the comparison to both datasets are presented and discussed.

Pg 6563, L24: State drawbacks of using aircraft data for this sort of general evaluation – for example the fact that you are using emissions which don’t match the year of observation. This would mean that for measurements gathered near sources, such as fires, which have a high inter-annual variability you may see large model biases.

We added the following statements at the end of Sect. 5.4: “The use of aircraft data for global model evaluation might have some limitations, due to the fact that model and observations are not always temporally co-located. This could imply, for example, that observations taken in the vicinity of strong emission sources (as biomass burning) could be affected by large temporal variability and indicate large biases when compared to model simulations.”

Pg 6565, L11 – 13: Maybe state the ‘positive bias’ is against the HadISST dataset, which is used to drive the TS2000 simulation.

Good point. We added that.

Pg 6565, L 12: Hard to compare inter-annual variability between levels as the plots are on different axis ranges. Have you calculated standard deviation weighted by the mean anywhere? This would give comparable values between levels.

We would like to keep the figures as they are. A normalized standard deviation would make the differences between the four simulations and the observations in each panel less clear. As we do not aim at comparing the results between different levels, we have not applied it.

Pg 6567, L11-14: Maybe mention that this is the upper limit of where nudging applies.

Done.

Pg 6567, L19: Change from ‘slightly weaker’ to ‘lower amplitude’?

Fixed.

Pg 6568, L1: Is there any particular region where these uncertainties are more common? High vs. low alts or tropics vs. extra-tropics? You can see from Fig 2 that the largest differences seem to occur in the upper troposphere in the tropics.

We modified this sentence as follows: “although there are some noticeable differences (especially near the tropical tropopause), revealing that uncertainties exist in the reanalyses as well”. Thanks for the suggestion.

Pg 6568, L2: reference the appendix where you describe Taylor diagram

We don’t think it is appropriate here, as this section discusses results. Methods are discussed in Sect. 4, and there Appendices A1 and A2 are referenced.

Pg 6569, L17-21: ‘The underestimation’ onwards - Slightly repetitive, combine sentences or shorten?

We rephrased this sentence as follows: “The underestimation of the west wind jets in the free-running simulations is an indication of an underestimation of the polar vortex. This is also supported by the warm bias in the seasonal mean of the temperature in this region discussed in Sect. 6.1.1 and shown in Fig. S2.”

Pg 6569, L28: Insert ‘observational datasets, particularly in the tropics’

Done.

Pg6570, L6: What do you mean by ‘simulation of the mean’?

We mean “the representation of the mean climate by the model”. We fixed this error.

Pg 6570, L17: Not sure I agree with ‘generally good agreement’. Fig S7 shows models are generally biased low at 500 hPa, 30hPa and 5 hPa.

That is true, but this bias is small in relative terms, only a few percent, as stated in the same sentence.

Pg 6570, L25: Why is 400 hPa more significant?

According to Gleckler et al., whose strategy is followed in our paper to calculate performance metrics, this is the most commonly analyzed level in the literature for this variable.

Pg 6570, L26: ‘In the extratropics,’ – ‘In the extratropics near the surface,’

We replaced this with: “In the extratropical troposphere”.

Pg 6571, L19: Where is the EMAC CERES comparison shown? If not shown then state ‘not shown’ in brackets.

Done.

Figure S11. Give units in caption.

Units are given in the figure.

Pg 6571: Can you say why the free running models tend to do worse? This is particularly evident in the long wave radiation comparison. Is this maybe something to do with the clouds that are calculated for the free running models? It would be useful to discuss this in a bit more detail at this point to give some idea of the uncertainty in calculating clouds for free running models.

One possible reason might be that the cloud and convective parameters have been optimised for the free running mode (see e.g. Mauritsen et al.: Tuning the climate of a global model, J. Adv. Model. Earth Syst., 2012, doi:10.1029/2012MS000154) and we kept them unchanged for the EVAL2 and

QCTM simulations. If the nudging systematically alters the cloud properties, the radiative balance will be altered as well.
We checked the total cloud cover and indeed found that the two free running experiments have a similar globally-averaged cloud cover (64%) which is higher than in EVAL2 (57%) and QCTM (60%). A sentence has been added in Sec. 6.1.4 to comment on this issue.

Pg 6572, L3. Need to mention what these sensitivity runs are previous to this point as they are shown in Fig 9 onwards. Put description in section 2.

We prefer to stay with the current structure, as the ACCMIP-S1 and ACCMIP-S2 are relevant for ozone representation, we think is better to introduce them in the context of ozone evaluation. Describing them in Sect. 2 would require to show the corresponding results for the climate parameters as well, which would be misleading.

Pg 6573, L15: What is the bias in the model in this SH peak ozone column?

We have quantified the differences and have added the following sentence to the text: “This positive bias ranges between 47 (EVAL2) and 59 DU (TS2000) compared to NIWA (49 to 61 DU compared to EOC-GOME).”

Fig 10: I think this can be moved to supplementary material. Also the lines are hard to differentiate between models.

We have moved this figure to the supplementary material as suggested.

Pg 6574-Pg 6575: Can you say something about the accuracy of total tropospheric column ozone from the MLS/OMI measurements? It would be good to know whether the model lies outside of this uncertainty range.

Ziemke et al. (2011) noted that using ozonesondes as reference, RMS uncertainties in local measurements of total column ozone from OMI/MLS are about 5 DU. They interpret differences of 10 DU and higher to be significant, while smaller values are essentially considered at noise level. We have added this to the text.

Pg 6576, L9-10: Not clear which model runs you are referring to here, give names of runs to be consistent.

Done.

Figure 15, Pg 6576: I think you need to make it clear in the figure that the ‘SH extratropics’ comparison only includes 2 stations and ‘SH tropics’ only includes 1. A table or map giving the station locations in the supplementary material would be useful.

The sentence on p. 6576 has been rewritten as it was incorrect. The same sentence has been added to the caption of Fig. 15 (now Fig. 16). A map has been added in the Supplement (Fig. S13) showing the location of the Tilmes ozonesondes stations.

Pg6578, L3: Add reference. Also mention that at very high concentrations of NOX ozone production become less efficient (so near source regions).

We expanded the text accordingly and provide an additional reference as follows: “At very high NO_x concentrations, ozone production becomes less efficient, because it is then limited by the abundance of NMHCs (Fowler et al., 2008)”

P6578-6579: I think the surface comparisons are of much better value then the aircraft comparisons as CO will have very high concentrations sampled by aircraft targeting emission sources such as biomass burning and anthropogenic emissions, switch the order of discussion (surface comparison first) and also mention the problem with using aircraft data climatology to evaluate models for a specific year.

Good suggestions. We reverse the order of the discussion (and Figs. 19 and 20) and mention again the limitations of aircraft data climatologies, which are also discussed in Sec. 5.4.

Pg 6579, L20-24. Models generally underestimate winter/spring tropospheric CO in the NH. This has been shown by Shindell et al., (2006); Shindell et al., (2008); Monks et al., (2014), ACPD (POLMIP multi-model comparison) against several other datasets. The reasons for this underestimate are still not fully understood.

We added this comment and the corresponding references. Thanks for the suggestion.

Pg 6579-6580: Several models have also been shown to underestimate ethane and propane in the northern hemisphere against surface data, see Emmons et al., (2014) in ACPD (POLMIP multi-model comparison).

Thanks for pointing us to this study. We added it to the text.

Pg 6582, L15-16: I think this CO change is due to a reduction in OH (you say this on line 11) so it is not going to influence the OH in return. However, the original OH change could increase the methane lifetime directly having important implications for climate.

We changed the wording to remove the circular argument. However, a more detailed discussion of the chemical effects of introducing the HO_2+NO = HNO_3 reaction is beyond the scope of this paper. Please refer to Cariolle et al., ACP 2008 and Gottschaldt et al., ACP 2013.

Pg 6583, L23: with exception of aviation emissions?

Corrected. Thanks for spotting this.

Technical corrections:

All the suggested technical corrections have been applied. Thanks for spotting them.