Response to the second 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.

This manuscript evaluates tropospheric composition simulations in the CAMS modelling system and quantifies uncertainties related to different chemical schemes. It is well structured and written and illustrates original and interesting results for the . CAMS modelling system. I suggest acceptance of the manuscript for publication after taking into consideration the following comments.

Main Comments

1) I guess that the simulations were carried for the year 2011 but I think the authors should describe in Section 2.3 which was the time period that the simulations were carried out.

Model simulations for 1 July 2010 to 1 January 2012 have been carried out. We now include such a statement explicitly:

To allow for sufficient model spinup, **the model versions are initialized for 1 July 2010 and ran through until 1 January 2012**.
The first 6 months of the simulation are considered as spin-up and therefore not evaluated.

2) The authors mention that the averaging of large number of measurements over space and time partly solves the problem of interannual variability (lines 273-275 in page 11). Can this dataset of Emmons et al. (2000) be representative to compare with the CAMS simulations for the year 2011? I understand the uniqueness of this dataset but could the authors clarify this issue and discuss the uncertainties and the weaknesses of this comparison?

Indeed the referee is fully correct with his analysis. It is also true that for the *total* anthropogenic VOCs emissions, the changes between the year 1990 and 2011 are of the order of 14%, following Emissions Database for Global Atmospheric Research (EDGARv4.3.2 database). Therefore obviously caution has to be taken when analyzing the comparison, but we believe the aircraft dataset comparison to be still a valid methodology despite the large temporal difference between observations and modelled data because of the following two reasons:
We expect the impact of this change to be lower at background locations or outflow regions, as included in the comparison of presented in this manuscript, only partly affected by anthropogenic emissions, while biogenic emissions are expected to remain largely unchanged. Also the variability and measurements uncertainties present in the observations are larger than 14%, implying that we can still consider these observations representative, especially because they are averages over larger regions in space and time. To make this clear, we now write in the manuscript:

For the total anthropogenic VOCs emissions the changes between the year 1990 and 2011 are of the order of 14%, following Emissions Database for Global Atmospheric Research (EDGARv4.3.2 database). Nevertheless, the evaluations presented here are all sampling background locations or outflow regions, and are hence only partly affected by such changes in anthropogenic emissions. Also the variability as well as measurement

uncertainties present in the observations are larger than 14%, implying that we can still consider these observations representative.

3) I would suggest the authors to provide a short description of the method used to calculate the weighted values of bias and correlation in Table 4.

The method has been extensively described in Jöckel et al. (2006), where the mathematical derivation is also explained in the appendix. As the mathematical description of the method would be too tedious, we have added the following short description to the manuscript, referring to Jöckel et al (2006) for detailed information. We now write

As explained in further detail by Jöckel et al. (2006), with this approach, the measurement locations with high variability have less weight, whereas more weight is given to stable, homogeneous conditions. This allows us to compare values that are more representative for the average conditions and to eliminate specific episodes that cannot be expected to be reproduced by the model.

4) The authors write in line 375 (page 15) that " CBA is the only model version to deliver a satisfactory bias" . Is this a robust conclusion? What is statistically satisfactory? Looking Table 4 I see that in some species CBA bias is smaller than in other schemes, in some other species the biases are comparable and in other species the CBA bias is worse.

Note that this specific comment on this line only referred to SO_2, so the reviewer is correct that biases in CBA are not always the best. In general, whether a bias is *satisfactory* small will depend on the application area, which is indeed not detailed in present work. Instead, in Sec. 5 we defined a weighted bias, which relates the bias to the standard deviations in the model and observations. This value should be between [-1,1] to deliver satisfactory results.
Therefore we choose to re-formulate the specific sentence here to:

For SO_2 CBA is the only model version to deliver a weighted bias that is larger than -1.

5) In the evaluation of ozone the authors conclude that "overall, the evaluation at individual station provides reasonable agreement between model simulations and sondes". How these evaluation results compare with other ozone evaluation studies which were based on MACC and CAMS products (e.g. Inness et al., 2015; Katragkou et al., 2015; Akritidis et al., 2018). I think this conclusion could be also supported by these studies.

So far we didn't show evaluation results from other CAMS products as this is beyond the scope of current manuscript. Furthermore, one should realize that chemistry versions and model configurations as adopted here are to some extent different compared to those use in important MACC/CAMS products such as the reanalyses.
However, to aid the interpretation of the model quality, i.e. to put the current model performance into perspective, we now include an assessment of the CAMS Interim Reanalysis (CAMSiRA, Flemming et al., 2017) in the evaluation against ozone sondes. We choose only to show figures for the annual average, zonal average bias and RMSE at various altitude ranges, to give a general indication of our model performance relative to that of CAMSiRA. Indeed, this evaluation shows that biases and RMSE are within the range of those of CAMSiRA, with the free running model versions of equal (or better) performance towards the boundary layer, and CAMSiRA generally better in the free troposphere. For further details about the configuration

and performance the reader is referred to Flemming et al. (2017) and Inness et al. (2019). We now write:

In this evaluation we also present data from the CAMS Interim Reanalysis (CAMSiRA) for the year 2011, to put the current model evaluation into perspective. This summary analysis shows averaged biases within ±10 ppbv, which is also in line with the O_3 bias statistics against the aircraft climatology. **At lower altitudes the model biases are mostly equal or better than those from CAMSiRA, while above 500 hPa CAMSiRA delivers mostly smaller biases thanks to the assimilation of satellite ozone observations.** The RMSE shows a larger spread in the lower troposphere of the NH, while at higher altitudes, above 500 hPa the overall magnitude of the RMSE for the three chemistry versions converges to values ranging from 10 to 16 ppbv, depending on the latitude. **Here the CAMSiRA shows overall better performance, mainly for the tropics and SH, while over the NH its performance is similar to IFS(CBA).**

6) In lines 448-449 (page 20) the authors write "Approximately half of the CO burden is directly emitted, and the rest formed through degradation of methane and other VOC's". Please add a relevant reference.

We now add Hooghiemstra et al. (ACP 2010) as reference for this statement, as they provide a detailed evaluation of a-priori and optimized budgets for global CO production.

7) On how many data points (and years) the temporal correlations shown in Figure 10 are based?

The temporal correlation presented in Figure 10 is based on twelve points (the monthly means) per station, and was evaluated for the year 2011. Therefore this figure shows an evaluation of the model ability to represent the seasonal cycle, as discussed in the manuscript.

8) In lines 526-528 (page 26) it is written "The vertical profiles (see Figure 13) are strongly biased (e.g., SONEX, Newfoundland and PEM-Tropics-A, Tahiti), with positive biases occurring at the surface and negative in the free troposphere."
Could this result also related to inadequate outflow from the atmospheric boundary layer (ABL) to the free troposphere (FT) and hence to model weakness in ABL-FT exchange?

Thank you for this interesting suggestion. Although such processes of vertical mixing could certainly contribute to uncertainties in the vertical distribution of C2H4, this would then also affect any of the other chemical tracers, as well as meteorological variables (e.g. humidity, temperature), which generally do not show indication of this type of issues. Therefore so far we have no indication that such an uncertainty is driving the discrepancy in the modeled vs observed C_2H_4 profiles, but rather believe that our emissions and chemistry contain larger uncertainties, as currently stated on the manuscript.

9) The authors refer to correlation R (that span from -1 to 1) but showing R2 which practically describes the explained variance. Although this is not crucial in the discussion it could propagate a misunderstanding on these statistical parameters when the article is published. I would suggest to modify this accordingly.

The reviewer is correct that there can be some confusion between the use of correlation in terms of R and R^2. Throughout the text we make sure to refer to R^2 when providing quantitative reference to the correlation, and at the start of Sec. 5 we now explicitly write:

“Table 4 summarizes the comparison of the various model results with aircraft measurements, described in Sec. 3.1, in terms of biases and correlation, **in terms of explained variance (R^2)**, …”

While or the table header we now write:

“Table 4. Summary of the Bias and correlation coefficients **(in terms of explained variance, R^2)** …”

10) Generally, I think that the discussion in model difference is rather technical and I would suggest the authors to discuss also the possible scientific reasons for discrepancies among the simulations with different chemical schemes for the different chemical species.

Indeed this manuscript has a largely technical focus, explaining the current state of the modeling system, its validation, and its general ability to quantify uncertainties due to model chemistry. Any more scientific reasons to explain discrepancies among simulations inherently require additional sensitivity studies, which is beyond the scope of this work. Having said this, we now do pay more attention to differences in model performance related to differences in their configurations, as also requested by the other reviewer, see particularly our responses to his ´Main Comments´. We hope this addresses the concerns raised here.

Minor comments
page 13, line 348: should rather be "relative shorter" instead of "relative short"

Thank you, we changed this accordingly.