Replies to Reviewer #1

General Comments: The paper by Dhomse and Chipperfield describes two new stratospheric data sets (TCOM-CH4 and TCOM-N2O), which were generated by a combination of TOMCAT model data and occultation measurements using a machine learning approach. These data sets are unique in a sense that different satellite instruments were used to generate the merged long-term stratospheric data set. The data are in general useful – as also written in the paper – for the evaluation of models and as a-priori information for retrievals. Both data sets are publicly available for download (has been checked). The paper is well written and contains all relevant information for data users except for some points addressed below.

We would like to thank the Reviewer #1 for his/her encouraging comments. Our replies are in blue italics. Briefly, we have done following changes (detailed response starts from page 2)

- *Clarified the zonal averaging procedure.*
- *Discussed the uncertainties of the measurement data and how they are considered in the method.*
- *Explained that the time dependence in the regression model is not very significant for most latitude bands.*
- *Defined the term "SWOOSH".*
- *Corrected the errors in the text regarding the vertical resolution of the ACE profiles and the averaging kernels.*
- *Corrected the sentence "which might introduce homogeneities".*
- *Confirmed that the feature importances in the XGBoost model are not normalized.*
- *Reworded a paragraph with a sentence "Hence attributing a single variable or a single processes is not possible".*
- *Added a summary of the results from Figures S5-S8.*
- *Explained the reason for the larger vertical variation of the TCOM-N2O profiles in the evaluation data set.*
- *Added some general information about Figures S9-S12.*
- *Deleted the sentence about the unusual data points in 2004.*
- *Clarified that the largest corrections are observed in the lower stratosphere for the absolute corrections, not the relative ones.*
- *Discussed issues with high altitudes / low concentrations measurements.*
- *Clarified the reason why seasonal shift in the atmospheric structure, low concentrations, high beta angles contribute to nosier retrieved profiles.*
- *Explained how only positive data points used for XGBoost training can affect the correction terms.*

I have the following general comments/questions:

1. The provided data sets are zonal averages. However, the zonal averaging is not explained in the method. Is it done before or after training/application of the correction term? Please clarify in the paper.

We apologize for the confusion. In the revised manuscript, we clarify that we first calculate 3D (longitude/latitude/height) profiles twice a day (1:30 AM and 1:30 PM) before calculating the zonal mean. We then obtain the daily mean by averaging the 1:30 AM and 1:30 PM profiles.

2. It seems that in the machine learning approach the occultation measurements are considered as 'truth' for the training (and also later for testing/validation). How are uncertainties of the measurement data considered? How large are these, and how do they compare to the differences seen during testing/validation? Retrieval errors are mentioned in the paper, but only in a general way without any quantitative assessment. It is also unclear, if/how the vertical resolution of the measurements / averaging kernels etc. is considered in the method. This should be addressed in the paper.

We thank the reviewer for pointing out the lack of information. In the revised manuscript we added a paragraph to explain that we consider the measurements with positive values and retrieval error less than 100% to be an absolute truth and our attempt is to construct the data that would approximate HALOE/ACE if the instruments had denser measurements without any temporal gaps. We also clarify that we do not consider averaging-kernel-related information (ACE does not have averaging kernels) as it is impossible to get similar information for all the model grid points.

3. How relevant is the time dependence in the regression model? Would it be possible to use this correction also for times not covered by the measurements (it seems so as the evaluation period is after the training period)? What do you think are the limitations?

##*Indeed,* the time (date) term is included in the XGBoost model to allow it to extrapolate corrections to data that lies outside the training period. However, in current setup, the feature importance of the time term is only significant at a few levels for some latitude bands. This suggests that the time term is not playing a major role in the model's predictions for these latitude bands. To improve model's performance, we also tried to increase number of trees, use Huber/quantile loss functions, but none of the changes helped to improve time term's significance. We have added discussion in a revised manuscript. In summary, in a current setup time (date) term is not very significant.

Specific Comments:

1. p. 5, l. 121: Please explain SWOOSH.

Done

2. p. 5, l. 130: Note: 1 km is probably the vertical sampling of the ACE profiles; the vertical resolution depends on the averaging kernels.

Reviewer #2 (Dr Boone) correctly pointed out that ACE does not use averaging kernels. The forward model used in V4.2 retrieval uses 1 km vertical resolution, hence fitted spectra are interpolated at 1 km resolution.

3. p. 5, l. 133: Limiting retrieval errors to <100% is a quite coarse filter. Since the correction method does not seem to consider measurement errors this may be an issue at least at higher altitudes.
Yes, at higher altitudes it can add some biases, but the median profiles seem to be close to the median profiles from observational data. As mentioned by Reviewer #2, it influences correction term estimates at high (low values throughout the year) and low altitudes (winter/spring time minima). We are aware that we cannot construct perfect data sets but our aim is to construct gap-free data set but similar to ACE/HALOE profiles.

4. p.5. l. 136: A 10-degree latitudinal overlap between the bins is allowed… From the definition of the bins, it seems that the overlap is 20 degrees? Furthermore, the explanation …to include possible extreme variations in the training data set is unclear – why is the variation in the non-extended bins not sufficient? As I understand, the following sentence Estimated differences for overlapping grids are averaged… actually does not refer to this step of the method (calculation of differences) but to later merging of the data in step 6. These sentences should be re-formulated to clarify the above.
We have expanded that discussion and clarified that we use 20-degree overlapping.

5. p. 5, eq. 1: Please specify what exactly is meant with 'time' in this equation. Is it the absolute measurement time (e.g. UTC time) or local time? In general, units should be specified for all quantities (e.g. does CH4 refer to number density or a volume mixing ratio?).
Done. Time is measurement date, and all the tracers are in volume mixing ratio units.

6. p. 6, l. 176: which might introduce homogeneities Do you mean inhomogeneities?
Corrected.

7. p. 7: Are the feature importances normalised or not?
No, these are directly from XGBoost.

8. p. 7, l. 196/197: It seems that these sentences (esp. the value 18 km) refers to CH4 - please clarify in the text.
Done.

9. p. 7, l 199/200: Hence attributing a single variable or a single processes is not possible. The formulation is unclear (attributing to what?). In fact a similar statement is at the end of this paragraph. Please reformulate.
Done.

10. p. 8, l. 234/235: Please add a bit of summary information about Figs. S5 to S8, e.g. if results are similar for the other latitude bins or not.
Done, again highlighting that biases are largest for SHmid and tropical latitude bands.

11. p. 8, l. 236ff and Fig. 3: The vertical variation of the TCOM-N2O profiles seems to be larger than in the original data (both observation and model), especially for the evaluation data set. Please explain.
We have revised the manuscript to highlight issue with the use of only positive values for especially for the regions where concentrations are very low (especially upper stratosphere and lower mesosphere). We have also noted the lack of an explanatory variable that accounts for the strong winter/springtime seasonal minima at polar latitudes as downward transport brings N2O-poor air from mesosphere to the stratosphere.

12. p. 8, l. 245: Also, please add some general information about Figs. S9 to S12.
#Done. Again, highlighting the larger biases in the SHmid and tropics.

13. p. 8, l. 253: What exactly is meant with unusual data points in 2004? Do you refer to the larger values at 40 km? Please clarify.
We have reviewed this issue once again and looks like those points are not unusual as we see similar features for other years. Therefore, we have decided to delete the sentence.

14. p. 9, l. 2: largest corrections are observed in the lower stratosphere This is the case for the absolute corrections, not the relative ones.
Yes, we revised the sentence to clarify it and added that those biases can be considered as systematic bias due to TOMCAT set up.

15. p. 9, l. 259ff: Especially regarding N2O, how large is the error of the measurements at high altitudes / low concentrations? How reliable are the measurements at high altitudes?
See replies to the earlier comments. Also, as correctly pointed by the reviewer and Reviewer #2, at lower concentrations and closer tangent heights at lower altitudes for high beta angle measurements means some ACE retrievals converge for negative values. However, the black data points shown in Figures 4 and 5 are the ones with positive retrieval values and retrieval errors less than 100%.

16. p. 9, l. 261–263: As the ACE-FTS retrieval algorithm uses multiple micro-windows, there may be a seasonal shift in averaging kernels causing fluctuations in the retrieved profiles. Why do multiple micro-windows cause a seasonal shift? Do you mean that different micro-windows are used during different seasons? Please clarify.
As explained by Reviewer #2, ACE does not use averaging kernels, but the forward model uses 1 km tangent height spacing and at lower concentrations (and high beta angles) they might be shifted close to each other. So, we have reworded those sentences as:
"As the ACE-FTS retrieval algorithm uses multiple micro-windows, a seasonal variation in vertical structure of the atmosphere means interpolated radiances would have very little variations when N2O/CH4 concentrations are low. Also, when concentrations of a gas low measured spectra would show very little change between two tangent heights, leading noisy profiles. Therefore, N2O (as well as CH4) profiles show large variability at variability increases when tangent heights get very close together. Additionally, as mixing ratio values get close to zero, retrieved values can be negative. Here, we use only positive data points for XGBoost training, so that correction terms used here might be positively biased, influencing seasonal cycle effects in CH4 and N2O concentrations."

17. p. 9, l. 263/264: As we use only positive data points for XGBoost training... Due to measurement uncertainties the occultation profiles may contain negative data points. If you only use positive data for training this might results in a bias. Please clarify.
Yes, Reviewer #2 also pointed out this issue. In the revised manuscript we added a discussion and mentioned that this might cause some positive biases in TCOM profiles.

18. p. 9, l. 274/275: Why do you show in Fig. 6 daily means for TCOM but monthly means for SPARC data? Wouldn't it be better to use for the comparisons in both cases the same averaging time interval? Please explain.
We agree with the reviewer. Our aim was to show that TCOM data is available on a daily frequency, but for a direct comparison, we agree that we should have shown monthly means. The updated Figure 6 includes monthly means.

19. p. 9, l. 274ff: What is the difference between the ACE-FTS CH4 data set used in this study and the corresponding SPARC data set? Please explain.
The main difference is that the SPARC data set uses ACE v3.6 data whereas here we use ACE v4.2 data. We aim to release TCOM 1.1 data that will use ACE v5.2 data and use both positive and negative values to avoid possible causes for the positive biases seen at higher latitudes and altitudes. We also note that SPARC data uses somewhat earlier versions of Aura-MLS (v4) and MIPAS (v422).

20. p. 10, l. 310ff: Only a suggestion: Maybe the comparisons between TOMCAT and TCOM should be described before the validation with independent data sets as they define the expected accuracy of the TCOM data.
We have briefly expanded the discussion about the differences between TCOM and TOMCAT in the text. We have also expanded the discussion of the observation-TCOM differences, especially for evaluation period but translating this to the expected accuracy cannot be justified statistically, so we have refrained from doing so.

21. p. 11, l. 338 and Fig. 8: for some years CH4 differences are clearly distinguishable. Please be more specific here. Do you mean the occasionally high values in the tropics? Actually, a lot of differences at 45 km seem to be below the lower range of the colour scale (-15%). How representative are relative values at high latitudes where concentrations are low? What is the reference for the relative values (TCOM or TOMCAT)?
We agree at higher altitudes absolute values are much smaller, hence percentage differences may not provide enough information. We have added a caution note in the revised manuscript. We also reiterate that we have a limited number of ACE profiles in the tropics which is reflected in smaller R2 values. We also added a sentence in the caption: "Differences are calculated as 200(TCOM-TOMCAT)/(TCOM+TOMCAT)" and altered the contour range to -30% to +30% so that larger differences are clearly distinguishable.*

22. p. 12, l. 370: A possible explanation would be strengthening of the stratospheric circulation... What do you want to explain here? A trend or a non-existent / negligible trend (as mentioned in the previous sentence)? Please clarify.
We have reworded the sentence to mention that positive trends in the tropospheric emission should increase stratospheric concentrations, but if it is compensated by the stratospheric/mesospheric losses then it would lead to much smaller trends in the stratospheric N2O.

Technical Corrections:

1. In general, please check the text for missing ‘the’ in the sentences.
Done.

2. p. 5, l. 125: occulatation → occultation
Done

3. p. 5, eq. 1: I suggest that instead of the written quantity names (like ‘temperature’) variables should be used in this equation. Note that a sequence of italic letters in an equation could be (formally) misinterpreted as products of single variables.
Done

4. p. 8, l. 229: to -0.05 → to -0.05 ppm.
Done

5. Figs. 4 and 5: Please specify the CH4 and N2O unit in the figure or the caption. 3
Done

6. Suggestion regarding the data sets: It would be good to have the unit of zonal mean CH4 and N2O not only in the global attributes of the data sets but also (or instead) in the attributes of the corresponding zonal mean variables.
We are sorry for the mistake. To avoid duplication of the data files, we aim to release v1.1 data that will extend until December 2022 with some minor updates such as using ACE v5.2 data and those files will have correct global and variable attributes.