Review of Guo et al., Heterogeneity and Chemical Reactivity of the Remote Troposphere defined by Aircraft Measurements

The manuscript presents the development of a gap-filled database of observations from the Atmospheric Tomography (ATom) aircraft campaign, designed to provide unbiased measurements of the chemical composition of air over the Atlantic and Pacific ocean basins over a large vertical fraction of the troposphere. The paper compares the observed concentrations and 24-hour averaged photochemical fluxes (ozone production, ozone destruction, methane destruction) calculated from six different global chemistry models constrained by the observations with the same quantities sampled from freely-running simulations of these same models. Of particular interest, the comparison of observations and models is done in a statistical sense and aims to address the issue of whether global chemical models run at the resolutions currently used (1 or 2 degrees in the horizontal) are sufficient to resolve the distribution of photochemical reactivity seen in the much higher spatial resolution aircraft observations. The authors find that models should be able to reproduce the distribution of photochemical reactivities, but also find significant biases in the concentration of NOx that results in biases of the distribution of, in particular, ozone production.

The ATom observations are a fantastic addition to the set of measurements of the chemical composition of the troposphere we have and the approach of statistically comparing observations and models allows us to advance past the facile comparisons of long-term means or requiring models put the right plume in the right place at the right time. While the approach and the results have tremendous promise, the organization of the material makes it very difficult for a first-time reader to make sense of it. For example there are initial references to RDS_R0, RDS_R1 and RDS_R2 (Line 153, Table 1) without any supporting explanation, forcing the reader to search through the Supplementary Information or be patient to find some discussion of these differences between these data streams in the Results section. There is almost no discussion of how the RDS is calculated except for a generic reference to the Supplementary Information. Taking some of the text from Section S.2 (starting at line 428 of the Supplementary Information) would help to improve the ability of the reader to understand the manuscript. And while many of the models that are used to calculated the RDS are well known, there is the use of F0AM which is a box model evidently, but with no other information, and which has been run using ‘MDS’ (Table 1), which I assume is the Model Data Stream? Other details that would
help in the interpretation of the results, such as whether the dates for the driving meteorology for the Chemical Transport Models (CTMs) and for nudging the Chemistry Climate Models (CCMs) match the ATom measurement dates, can only be inferred. Many of the minor comments are directly related to the problems of organization of the presentation and, while individually these are rather nit-picky concerns, the cumulative effect on the reader is disorienting and it takes considerable effort and searching to understand the material being presented.

My other significant concern is with one of the fundamental results of the paper – that the spatial variability in reactivity calculated from the original ATom data should be resolvable by our current global CTMs and CCMs. The results are discussed in lines 345 – 353 and shown in Figure S8. Perhaps it is a problem of my understanding as the approach to calculating the results for Figure S8 is not explicitly stated: I assume you took random points from the P-O3 frequency distribution and then averaged the 10s data points that were adjacent in space and time along the ATom flight path? If the length scales inherent in the data were of the order 100 km, and thus resolvable by models, then I would think the frequency distribution would not change very much as three or five points adjacent in space are averaged. But Figure S8 shows a rapid change of the distribution towards a Gaussian centered on the mean of the original 10s data. So I am not able to understand how the results shown in Figure S8 support the idea that models should be able to resolve the spatial scales found in the ATom-derived reactivity. The results from averaging 8 adjacent data points (~16 km) shows almost no occurrence of P-O3 greater than 4 ppb/day.

Minor Comments:
Lines 144 – 156: This section has a discussion of the need for gap filling and introduces Reactivity Data Stream RDS_R0. There is also reference to Table 1 where RDS_R0, RDS_R1 and RDS_R2 appear. But the reader must dig into the Supplementary Information for any idea of what RDS_R0, etc. refer to. There is not even mention that the nomenclature refers to different approaches to deal with data gap filling. It is in the Results section, starting at line 167, that there is some discussion of R0, R1 and R2. The history of the different MDS versions and how they led to different RDS versions needs to be coherently introduced.

Lines 151 – 162: This section describes the RDS, but there are no details on how the RDS is calculated, forcing the reader to go to the Supplementary Information or back to Prather et al. (2017). The text of this manuscript should include sufficient details to allow the reader to make sense of the article as they read through it so I would urge the authors to include some minimal outline of how the RDS is calculated by the different models using the MDS as input.

Lines 164 – 165: The reader is referred to Methods for information on how the MDS was constructed. There does not appear to be a Methods section in the body of the article. Do the authors mean to refer to the Supplementary Information? For that matter, there is also reference to a Methods section at lines 113 and 116.

Lines 186 – 187: ‘We include the statistics from UCI using alternate years (1997 and 2015 versus the standard 2016) to show the effect of different cloud fields.’ underlines the lack of details in the manuscript about how RDS is calculated. It is not mentioned anywhere that I can find what years were used for the RDS calculations.

Lines 228 – 231: The authors find that the model calculations of RDS show quite similar distributions of reactivities when constrained by the MDS (Figure 1). How does meteorological variability fit into this comparison. There is not much information on how the RDS was calculated – time and space matched to the ATOM flights using CTMs and nudged GCMs, sampling a number of different years or just a single year – so it is difficult
Lines 293 – 294: ‘The complex patterns of the 3Rs seen in Figure 2 cannot be matched directly with CCMs’. From Table 1, at least two of the CCMs were nudged to reanalysis. Would that not provide a similar level of fidelity for transport and airmass history as the CTMs? Later, at lines 318-319, there is the mention ‘this could be tested with CTMs using 2016 meteorology and wildfires.’ so it seems even the CTMs were not run with meteorology specific to the ATom campaigns? This is all well and good, but another example of the way in which a very ‘thin’ description of the setup makes it very difficult to interpret the results.

Line 303: A duplicate ‘that’ in ‘indicate that that P-O3’

Lines 328 – 344: This paragraph discusses Figure S7 that is found in the Supplementary Information. I would suggest moving Figure S7 to the main body of the article if you are going to discuss it at any length.

Lines 345 – 353: As for Figure S7, I would suggest Figure S8 move from the Supplementary Information to the main body of the article.

Lines 372 – 375: On the disagreement for HOOH (‘If anything, the models tend to have too much HOOH: ATom shows systematically large occurrences of low HOOH (50-200 ppt, especially Central Pacific) indicating, perhaps, that convective or cloud scavenging of HOOH is more effective than is modeled.’) I agree the scavenging could certainly be the source of the problem. And, while it is equally speculative, I can’t resist pointing out that an overestimate of HOOH photochemical production would agree with the low bias for NOx found in the models. Is there any correlation between the regions where HOOH is overestimated and NOx is underestimated?

Line 615: I was not able to find any captions for the three tables.

Line 615: In Table 1, two of the three CCMs explicitly mention nudging for meteorology but NCAR (CAM4-Chem) just says ‘MERRA’. Was it also nudged to MERRA?

Supplementary Information

Lines 109 – 110: It takes digging into Table S2 to deduce that the NOx (PSS) calculated for MDS_R0 seems to refer to the calculation of the NO2 concentration from measured NO and assuming PSS. Starting at line 194 we learn a little bit more about the problem with NO2, but it is still not quite clear how NOx for the the final MDS_R2 was calculated. Were these data points dropped? The description of this problem, in particular, should be a little friendlier to the reader who is coming to this data for the first time.

Line 258: What is CO_N in ‘Create a continuous CO_N record.’?