The authors appreciate the time and effort of the reviewers in providing constructive criticisms that have helped us improve the paper, making it more readable and useful to the community. We address most of the issues with some minor revisions such as the inclusion of more explanatory supplemental material as documented in our responses below. Two more significant changes were deemed necessary.

First, in response to the reviews we added a ***Prologue***. Given the incredible amount of detail in the two preceding papers (Prather et al., 2017 ACP; 2018 AMT) that must be understood to read this paper, we chosen the unusual approach of adding a ***Prologue*** that summarizes the key results from those papers that are relevant to this one.

Second, we found an inconsistency between the reported concentrations of both pernitric acid (HNO_4) and peroxyacetylnitrate (PAN) and the chemical kinetics used in the models while investigating sensitivities and uncertainties in the ATom chemical species for future studies. High concentrations (attributed to instrument noise) were reported under conditions where the thermal decomposition frequency was >1 per hour. There is no easy fix for this, and we left the species data in the MDS as it was reported, but damped the two concentrations based on their thermal decomposition frequencies before the reactivity calculation. This is effectively a new protocol RDS* for calculating the reactivities, see Table 2 and description in revised text. There was also some confusion on how and when we applied the cosine(latitude) weighting, and this is now corrected. These corrections resulted in the tabulated numbers shifting a few percent in some cases, and in marginal shifts (often not detectable) in the updated figures. We now use the RDS* protocol for the reactivities calculated from MDS-2 in the figures, and both RDS and RDS* results are shown side-by-side in the tables.

We believe that this revised manuscript is now clearer. The original approach for analyzing aircraft and model data did not change, nor did any of the major findings. We hope it is now approachable and useful to the community, and can proceed to publication in ACP.

Reviewer #1

> *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.*

Thanks for the careful read, yes, this is what we intended.

> *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?

We concur. The current manuscript is difficult to follow. To start with, we add a Prologue Section before the Introduction Section that summarizes the key results from the previous two papers. The detailed protocols and first-order results from the pre-ATom papers (P2017, P2018) may be difficult to extract from these papers and the Prologue provides an important introduction to this paper. In addition we modified/rearranged the models, methods, and data sections to merge more of the material into the primary text. More detailed description for MDS and RDS versions were added and noted throughout the text. Captions for tables and figures were augmented to avoid abbreviations or telegraphic text. Further, we changed the notation for the MDS versions to be simpler: MDS-0, MDS-1, MDS-2. We have been stuck with annoyingly cumbersome notation of the aircraft mission data archive (insisting on the '_R0, _R1, …' suffixes to each data set, no matter how minor the corrections), but can streamline this for the paper. Also we note that the RDS has had a single protocol (method) for the calculation until this revision, and RDS-2 now means RDS using MDS-2. As explained above, we needed to introduce a new protocol for future work and that it simply denoted RDS*.

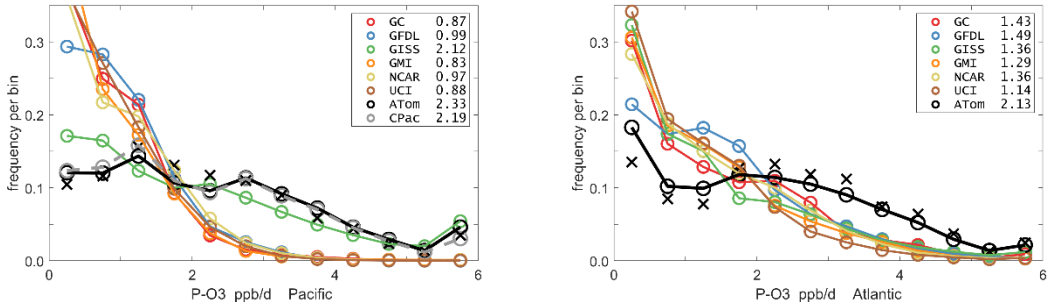
Other details that would help in the interpretation of the results, such as whether the dates for the driving meteorology for the Chemical Transport Models (CTMs) and for nudging the Chemistry Climate Models (CCMs) match the ATom measurement dates, can only be inferred. Many of the minor comments are 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.

We now explain in the prologue that our long-term goals were to establish a chemical climatology metric that focused on the budgets (i.e., reactivities) of ozone and methane, and which could be applied to the climate (free-running) CCMs, since nudged models are always different (issues about convergences, cloud systems, etc.).

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.

We thought this was an interesting result, and indeed it may be fundamental. This idea pushed us to examine the 1s versus 10s data where we had good 1 Hz measurements (H_2O and O_3 as shown in Figure S2). That case study was encouraging, and so we tried to assess what would happen with ATom PD (Probability Density) for something like P-O3 over an ocean basin if there were high-frequency variability below the model grid scale. This was Figure S8, which we clearly failed to convey that it was only didactic. So we have found a more convincing case and dropped Figure S8 discussion entirely.

Per other review comments, we moved Figure S7 (comparison of 1D PDs for the 3 reactivities from ATom and from 6 model climatologies) into the main text as Figure 5(new, one panel below). This Figure 5 now includes ATom PDs derived from the gridded data plotted in Figure 2 (plotted as black X's). Figure 2 has the ATom 10s parcels data averaged into 1-degree latitude by 200 m altitude blocks, effectively the resolution of the best current global models. Thus, when we sample the ATom 10s reactivities (O's) on a model grid scale (X's) the PD shifts only slightly (yes, we must lose some of the variability, and we see it), but this shift is less that the differences between models, and thus we can use aircraft PD statistics (at least over the ocean basins) to test the chemical reactivities (and species) in the CCMs.



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.

Yes, indeed. We have expanded the main text (reducing the supplemental) and put the discussion of the versions in a single place. We also have simplified the notation: MDS-0, MDS-1, MDS-2. We note that the RDS has had a single protocol (method) for the calculation until this revision, and RDS-2 now means RDS using MDS-2. As explained above, we needed to introduce a new protocol RDS* for future work.

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.

We agree. This problem should be solved with the new Prologue.

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.

Sorry, that was a bad mistake. We pulled more of the methods and supplement into the main text and correctly labeled the 'Supplementary Information'.

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.

Good point. This problem should be solved with the new Prologue and other text revisions to the description of RDS. With UCI (we did not burden the other contributing models on this) we calculated the RDS based on the year of deployment (2016, not all models could) and we calculated alternate years

(1997 and 2015) to show the effect of different cloud fields. The other models' years vary and most were not specific meteorology for that year but fixed SSTs for the CCMs.

> *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 to judge.*

With the new Prologue we hope it is now clear that the meteorological variability does NOT affect the RDS calculated by the models, ***except*** for photolysis = clouds, which come from the models. That is why we did an earlier assessment of model J-values (Hall et al., 2018) and include J-value diagnostics from the models in Table S8. We have a discussion in the earlier papers (P2017) that there are no accurate assimilated cloud products at the level that would allow J-value calculations, e.g., the ECMWF and MERRA-2 fields will have similar frontal systems and tracer transport, but the cloud systems are not consistent. The RDS requires us to estimate the J-value history for each parcel. Hence we accepted that as an inherent uncertainty and asked the models to average over 5 synoptically separated days. The different years of the UCI-CTM (1997, 2015, 2016) gives us an estimate of that uncertainty.

> *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.*

We believe the new Prologue has laid the groundwork to understand why this approach will not work. We prefer not to belabor it in the paper. For one we need to follow the parcel for 24 hours, and must make the approximation that it does not move or mix with neighboring air (which of course it does, but cannot be defined from the aircraft data). Also, a longer discussion, not here, is that exact matching of frontal boundaries – even with a good CTM – never quite works. We tried it with TRACE-P and gave up. Nudged CCMs do not really do the same job as CTMs (always some internal dynamics) and the nudged CCM has inherently different mean properties (clouds, convection) than the free-running CCM.

> *Line 303: A duplicate 'that' in 'indicate that that P-O3'*

Thanks, typo is fixed now.

> *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.*

Yes! That is an important figure and is now moved to the main text as Figure 5.

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

We have given up on Figure S8 and its analysis (described in detail above).

> *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?*

You have a good point: Could the model error toward high HOOH cause an error toward lower NOx? It seems reasonable. We would have to do some additional model sensitivity tests (next paper). In terms of looking at the correlation of HOOH and NOx, we have that in Figure S6(old). The 2D PDs seem fairly symmetric with the NOx peak distribution occurring in the middle of the HOOH distribution, no obvious correlation. We are barely able to compare statistics from the ATom transect vs the Pacific Basin of the models; I do not think we can separate out regions. In view of the possible importance of this figure relative to Figure 4 (the 1D PDs) we have moved it into the main text as Figure 7(new).

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

Oops, we have added captions (extended titles) for each table and included notes where appropriate. Our habit is not to add captions to tables, but use "Notes:" if needed.

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?

Correct, this is now fixed to: '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 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.

Yes, good point. We have expanded the text to explain the sequence of MDS versions, how MDS-0 uses NO and O_3 and J's to estimate NO_2 and thence NOx. In all later versions (MDS-1, -2), we use the observed NO_2 to calculate NOx. We added a briefer description of the MDS versions in main text

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

We have added explanatory notes at several points in the text here. The continuous CO record was a holdover from MDS-0 and -1 and used to fill gaps in the MDS record using CO as a proxy for variability across gaps. We found after building this apparatus, that the CO did not seem to correlate very well with other species and dropped that approach for gap-filling.

In the section on P & T:

"In this document we are careful to give measured species a suffix that denotes their provenance, and thus the variables denoting the combined, continuous data are P_M and T_M."

In the section on CO:

"CO. In our first attempts to produce a gap-filled record for chemical modeling (the MDS), we sought a species with continuous measurement that could be used as a proxy for unusual or polluted air during the gaps in other species. CO was the obvious species because it is indicative of biomass burning or industrial pollution and ATom has two well calibrated, nearly continuous measurements: CO_NOAA and CO_QCLS. The primary CO data are from QCLS because it has higher precision and the secondary are from NOAA which has fewer gaps. Unfortunately, after creating this gap-filled CO data and applying it as a proxy, we found that CO had little skill in filling the gaps in other species. We use this method to generate our CO_M record for MDS-2, but do not use it for other species in MDS-2."

Reviewer #2

This manuscript presents (1) several gap-filling methods for the ATOM dataset, creating a model data stream (MDS) (2) methods for calculating reactivity data streams (RDS) using the MDS, and RDS for six models, (3) a comparison of modeled and measured reactivity, with a specific emphasis on the ability of coarse resolution models to capture observed spatial heterogeneity. The authors conclude that models are capable of reproducing the statistical distribution of measured reactivities.

Yes, that is a good description. One of our results was not that the coarse models capture all the spatial variability but rather that they produce statistics of the chemistry, including reactivity of air parcels, that capture much of that observed variability at 2 km scales. We have been able to strengthen that case with added analysis of the data in Figure 2 to generate PDs for Figure 5(new), please see response to Reviewer #1 above.

The work closely follows the analysis in Prather et al. (2018), and adds a powerful observational dataset. The MDS and RDS datasets are valuable to the wider atmospheric chemistry community, and the approach has a solid foundation. Further clarification of methods and a more careful quantitative analysis would greatly improve the manuscript. Comments below refer both to content and clarity.

Major:

It would be helpful if the key reactivities were defined early in the manuscript rather than in the supplement.

Yes, very good point. We have added the Prologue Section and moved the description for RDS from SI to main text.

The F0AM model can be configured with a number of chemical mechanisms (i.e. F0AM itself does not have reactions, but relies on the MCM, GEOS-Chem, Carbon-Bond, etc). More specification on the F0AM setup is needed. For example, in lines 215-216, the authors refer to the "F0AM protocol for NOx"—where does this protocol come from? It seems the box model could be set up such that NOx can photochemically evolve. The rational for discrepancy in model procedures need further explanation.

The description of F0AM chemistry (MCMv331 plus photolysis of HNO_4 and CH_4+O(^{1}D)) and their protocol (let NOx decay over 24 hours like the A-run) for these results is now corrected.

It is unclear why the RDS_R0 is used when it contains known errors, that were later fixed (lines 171-172). It seems using the most accurate RDS is possible (lines 180-181), and it would yield the most useful paper.

The MDS-0 has known biases but it is not that different in terms of the range of results. We asked the six models to run their RDS calculation of reactivities using MDS-0, and it took a while to collect those results (RDS-0). The comparison across models with RDS-0 is adequate to assess the differences and similarities across the models, and that is what they are used for here. We did not wish to burden our community with unnecessary model simulations, as they take up a lot of people-time setting up each one (because the MDS structure changed). When we switched to observed NOx with MDS-1, we compared GMI and UCI and found the same level of differences as for MDS-0. When we then corrected the gap-filling errors to reach MDS-2, we only ran the UCI model. The shifts in reactivity statistics are shown in the tables here. With this paper we are publishing the MDS-2 for all 4 ATom deployments, and that is the dataset we shall analyze with as many models as possible.

With this revision, we add a refinement to the RDS protocol (RDS*) by damping the concentrations of HNO_4 and PAN (no new MDS are generated) according to their thermal decomposition lifetimes (denoted UCI2* = UCI model using RDS* on MDS-2). This revision in the RDS methodology causes barely discernible shifts (~3% or less) in the figures or model statistics shown here, but greatly affects the sensitivity calculations being done for a subsequent study.

One of the supporting pieces of evidence that that models can capture spatial heterogeneity is the descent

given in the Supplement Figure 2. The analysis of this is purely visual. It looks as if some of the reactivities may vary by up to 50% in a given 500 m box. What is the variability? What would be considered an "acceptable" level of heterogeneity in a box?

Yes, indeed Figure S2 is only visual evidence, and only a case study. It is relevant to show that scales of variability go from 1 s to 10 s to 100 s. It is not used as proof. Figure 5(new) includes a robust statistical view of the impact of the fine-scale structures on the reactivity PDs. We do not know what "acceptable" means here, there will certainly be some example of fine smoke plumes with sub-10s structures.

Line 287 says "the spatial scales of variability are within the capability of modern global models", but directly after, line 293 says "the complex patterns of the 3Rs seen in Figure 2 cannot be matched directly with CCMs". It seems these two statements are incompatible. Can you clarify?

Sorry for the confusion here. Figure 2 shows reactivity structure at the pixel level (100 km x 200 m) that is resolvable by the 1° models, but even if the model grid cells are this size, these models are incapable of simulating the ATom flight with any accuracy. Forecast, assimilated, or nudged models cannot match frontal passages or convection at the level of these observations (10 s and 2 km), and so we do not expect any model to produce the same exact "hot spots" in reactivity, but in a statistical probability density, yes. We modified the statement to "cannot be compared directly with CCMs". That is why also assess mean profiles of reactivity in Figure 3.

Paragraph starting at line 345: It is unclear to me how this analysis and Figure S8 supports the conclusion that "the ability to nearly match Atom-statistics is […] significant" (also, what is meant by a significant ability?). Perhaps a dummy argument would help.

Figure S8 was created as a didactic figure that did not work and has been removed in favor of the new analysis shown in Figure S7(old) now Figure 5(new).

Paragraph starting at line 354: The authors lead the reader to assume the models are missing a lightning NOx source. This would be a major conclusion that needs to be placed in the context of other literature. But also, I am wondering about the discrepancy in NOx between MDS versions discussed earlier in the manuscript. Has that impacted this analysis? The same comment applies to the following paragraph. Clarifying notation and using the most accurate datasets would help readers.

We will reword that paragraph to make clear the discrepancy: The model results are from the free-running model climatologies for August. All the models have lightning NOx and are not using any MDS data in this case. In this case and figure, the ATom data is the UCI CTM in A-run mode (RDS* protocol) using MDS-2 (observed NOx). One plausible explanation is missing ocean emissions of NOx or NOx precursors, but that requires the NOx to reach to 4 km (difficult for surface source) and ATom did not identify any obvious NOx precursors. The hypothetical proposal is that the redistribution of lightning NOx may be wrong, and more should from oceanic lightning and be redistributed to the lower tropospere.

The conclusions sections presents two new figures (figure 5, figure s6) and two new "quick look" interpretations. These brief analysis are cursory, and not conclusions of the paper.

Correct, they do not contribute to conclusions, but rather to paths going forward with the type of analysis shown here. Thus we have renamed the final section as Discussion and Path Forward.

Minor:

Line 92: "most models agree in the CH4 and O3 chemical budgets": does this mean "terms in budgets" or budgets themselves?

Thanks, our mistake. We meant "on the budgets".

Line 170: Comments like "Three central models showed excellent agreement" are vague. What agreed? How do you quantify that agreement?

In Table 3, the cross-model RMS showed 20-30% differences between 3 core models. We modified the statement as "Three central models showed excellent agreement as shown in Table 3". Their cross-model

RMSD were only about twice that of the RMSD for the same model using different years for clouds and photolysis.

Line 183: "In our analysis, the ATom 10s parcel s are weighted to achieve uniform sampling": What does this mean? Is there some post-processing weighting of the observations?

We did not change any of the measurements in post-processing, but parcels can and should be weighted differently when calculating mean values or probability densities "to achieve uniform sampling densities." For example, if we want a tropospheric average, then we must recognize that there is more flight data at cruise level than in the marine boundary layer. We have expanded and made more explicit the discussion of weighting.

Line 197: "Key photolysis rates are similar across all model except GISS, and because of this and other inexplicable results…" Should we assume that GISS model is fundamentally different than the others, or that there was some unexplained error in the model setup?

GISS model results remain peculiar, inexplicably different from the other models. We cannot determine if this is inherent to the model or caused by the model setup for these simulations. However, this problem extends even to the standard August climatology in P2017. One of the co-authors has tried to identify possible problems but has not yet found them. So "fundamentally different" is all we can say for now.