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.

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.
- 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.
- 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.
- 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?
- 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?

- 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.
- Paragraph starting at line 354: The authors lead the reader to assume the models are missing a lighting 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.
- 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.

Minor:

- Line 92: “most models agree in the CH4 and O3 chemical budgets”: does this mean “terms in budgets” or budgets themselves?
- Line 170: Comments like “Three central models showed excellent agreement” are vague. What agreed? How do you quantify that agreement?
- 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?
- 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?

References: