Prather et al report a new modelling protocol for merging in situ measurements of reactive chemical species into 3D models that simulate their chemistry. The goal of this work is to enable new insight into how the chemistry schemes and processes in these 3D models affect the results of simulations performed with them by isolating (switching of) meteorological mixing, emissions and deposition. This protocol is tested with a synthetic (model derived) data stream in the current work but could be extended to a real observational data set when such a data set becomes available (i.e. from the ATom mission dataset).

This is a very original idea – albeit an extension, in some way, of years of previous work using single box models. It is a neat idea and as a global modeller not involved in this work I will be working out how to apply this protocol to my model to test my model's results! I would recommend publication in AMT following the adoption of the following minor points.

My major annoyance with this paper was the constant reference in the main paper to figures and tables in the supporting information (SI). I would recommend large parts of the SI be moved into the main manuscript.

*OK, have moved many of these into the main paper. In particular, we have brought all the 6-model plots into the main paper, but left the one-model sensitivity test figures in the Supplement.*

General comment:
• Many of the figures are low resolution. Please make sure that the final images are higher quality scalable vector images. In addition, many of the figures have no units on the axes. Please check them and add units where appropriate.

*The .png figures in the .doc file look sharp but my conversions to .pdf made them fuzzy – sorry, should have looked more closely. In the final, all figures are fully resolved .eps*

• Please also ensure consistent use of colours for models. There are many figures and it makes it easier to keep track if colour is used consistently. For example in Figure 2 NCAR is black but Figure S2 NCAR is magenta.

*Agreed. I am still learning about how to use colors with matlab and will implement better, consistent colors!*

Specific comments:
Page 1, line 21: "the data" I think should be "these data".

*done, thanks.*

Page 2, line 36: Are the regions being referred to spatial or chemical or both?

*done, 'spatial'.*

Page 2, line 60: Please add the information referenced in the SI on Reactivities to the

main document.

*done.*

Page 3, line 72: How large is large? Can this be quantified?

*Miswrote here, the subject is 'causes' and the predicate is now 'are many.'*

Page 3, line 75: This is a minor point but as I understand PhotoComp 2010 mainly assessed stratospheric chemistry? The work in this paper focuses on tropospheric reactivities. What relevance is this reference to the present work?

*Actually PhotoComp performed a range of photolysis, including J-NO2 with tropospheric cloud layers.*

Page 4, line 118: I think this refers to Tables S1a and S1b not S1?

*done, thanks.*

Page 5, line 131: I think it's important that the information referred to here from the SI be put into the main paper. This is a technique paper and it is painful to have to refer to the SI to actually find out about the technique in question.

*OK, OK. Have expanded:*
*"The implementation of this data stream of reactive species is model dependent.  All models begin with their own 3D initialization data set that is used to restart a model simulation beginning on August 16.  The specified air-parcel NOx, for example, will be initialized as separate NO and NO2 abundances by scaling the model's restart values for NO and NO2 to match the specified parcel NOx.  Similarly, a single value for aromatics will be partitioned over benzene, toluene, and xylene by models that resolve these species in accord with the restart values.  The models place each parcel (i.e., overwrite the restart values) in the grid cell containing the latitude, longitude, and pressure specified for that parcel.  If that preferred grid cell is already occupied with an air parcel, then an alternate adjacent grid cell is selected.  It is recommended that alternate cells be shifted to minimize the change in photolytic environment (e.g., shift by longitude but maintain surface albedo and atmospheric mass).  Two chemistry-climate models (GFDL, NCAR) were unable to completely overwrite the modeled T&q values with data stream values (see sensitivity tests below).  See also Supplement for additional details."*

Page 6, line 168: Define RMS.

*done, thanks.*

Page 6, line 168: It's not clear why 3 different model years would be capable of defining a lower-limit for the RMS? Why not 2, 5 or 10 years?

*The choice of 3 seemed reasonable to test 3 different pairings. The sentence have been revised for clarity:  "UCI ran 3 different model years to estimate the rms value caused by interannual variability (blue in Table S2), i.e., when the cross-model differences approach this value, we can accept that the photochemical modules including clouds cannot be said to be different in this study."*

Page 6, line 168: I'd suggest adding "(in blue in Table S2)" after "lower-limit rms".

*done, but part of revised sentence above.*

Page 7, line 206: Are there any references for this statement of "ambiguous choices"? Will other modelling groups know what you mean here. I'm afraid I don't.

*Sorry, I thought this was well known, but will be explicit and add: "(i.e., the absorption cross sections are given at 220K and 294K; the quantum yields, at 248K and 298K; and the choice of whether to interpolate linearly or logarithmically, or whether to extrapolate or not, affects J-NO2, especially in the upper troposphere). "*

Page 12, line 280: Correct the duplicate "that" typo.

*done.*

Page 13, line 315: From Figure S6 it seems to me that GISS is an outlier for L-CH4. Can you clarify why you state it is not?

*Sorry, this appears to have been written with an earlier set of figures, it requires a major re-write: "Focusing on the upper 10%, the outliers are unusual and reactivity specific: for L-O3, GFDL is much steeper that the other models, consistent with the feature identified earlier in the scatter plots; and for L-CH4, GISS is much shallower. Surprisingly, with this diagnostic GISS is not an obvious outlier for P-O3 and L-O3 as seen in previous comparisons."*

Page 15, line 376: Typo – "great" should be "greater".

*done, thanks.*

Page 16, line 378: I would suggest adding "(solid line)" after "reference case".

*done, thanks.*

Page 16, line 386: Looking at Figure S2 I don't see that the largest J-NO2 values are within clouds. It looks to me that J-NO2 follows J-O1D and increases to a maximum in the upper troposphere.

*Correct. Fig S2, because it is averages, and does not show the cloud levels. I will make the statement because it does explain the patterns and does follow from our (not shown) studies.*

Page 16, Figure 3: Please add x and y axis information to the plots.

*done, thanks.*

Page 17, Figure 4: I like the figure! But, the colours may be difficult for a colour blind person to interpret. I have had a go at uploading and examining the figure here (http://www.color-blindness.com/coblis-color-blindness-simulator/). As far as I can see this may be OK, but I would advise as a general note to consider the use of colour more.

*Good point, but not really necessary as the colors are the same for all the Venn diagrams and only used to guide the eye. I will see how easy it is to redo the colors.*

Page 17, line 398: Why is this surprising?

*Good point, dropped.*

Page 17, line 403: Remove typo "models".
*done, thanks.*

Page 18, line 412: Please clarify that 8/16 is 16th August. As a non-American English user I was a bit confused but guessed this is what you meant?
*OK, put "August 16"*

Page 18, line 419: Are these small differences likely in all environments? Is this protocol only valid for the remote Pacific? What levels of NO3 and N2O are present and how would a more vigorous night-time chemistry effect these conclusions?
*I really do not want to get into a lot of 'what ifs' here. A caveat sentence has been added.*

Page 18, line 435: Delete "out".
*done, thanks.*

Page 18, line 437: I think "J-values" should be "J-value".
*done, thanks.*

Page 19, line 440: I think removing "5x14,880" and replacing with "5 days that we consider 14,880 parcels per day," would make this a bit clearer for the reader.
*Yes, fixed another way.*

Page 20, Figure 6: What is "Ud4"? in the legend? Delete "use of their" from the caption.
*OK, need to rewrite the caption for clarity.  Ud4 is UCI running day 4 (Aug 16) with the GFDL or NCAR T's and T&q's.*

Page 21, line 495: Decapitalise "Overall".
*done, thanks.*

SM-1: There is a typo in the units for P-O3.
*done, thanks.*

SM-2: It is not clear how you have adopted the partitioning of collective species?
For example, in my model I only have 3 NMVOC. My ATom or synthetic data set has 12 NMVOC. What should I do?
*Good point, I have added: " The algorithm for dealing with missing species or an over-specified class of species is truly model dependent.  For example, the UCI model has a simple approximation and single class for all aromatics and consolidates emissions of benzene, toluene, and xylene into 'aromatics' that react as benzene.  The NCAR model includes all three species explicitly, and thus they will take the mole fraction of 'aromatics' and partition it into benzene, toluene, and xylene, scaled to their values in the grid cell that is being overwritten with the UCI data stream. "*

I think it would be nice if the UCI data were made available as a data source for other modellers to use and then they can compare their results to this paper.

*Humble apologies, yes, of course we will publish the data stream used here with the SM.*

SM-4, Table S1b: Can the versions of codes to calculate photolysis rates be added? Similarly, can references for the chemistry schemes or the papers that describe the schemes be added. I'm sure GFDL have updated some of the rates that were in the original MOZART-2 scheme since it was developed. ASAD, as I understand is not so much a chemical mechanism but a software for integrating chemistry. What is meant by "Wild, FRSGC"? Is this a reference?

*I will try to extract some more details, but a full documentation of the chemistry is beyond this work. The UCI notes-to-self that you noted have been replaced with a proper reference.*

Table S6, S7 and S8: What is "R"?

*Sorry, that was lazy, have spelled out "reactivity"*

Figure S1: There is a growing consensus that the rainbow/jet colour scale should not be used for quantitative inference (Hawkins et al., 2015). I would consider switching the colour scale used to something like the viridis colour scale.

*A reasonable point, but this figure is not primary, and will leave as is for now. Nevertheless, I am reworking the colors and legends for the other key figures.*

Figure S1-S9: See general comment about consistent use of colour. NB I can not see a difference in colour between GFDL and GISS in the legend of Figure S8.

*Yes, see above. We are moving S2, S6, and S9 into the main text.*

Figure S11: What is "Ud4"?

*Yes, revised caption and figure. No need to specifically call out UCI day 4 (8/16).*

Figure S12: Are the different dates important here? If not a hex-bin plot would be much nicer to help the reader see the relationships between the data.

*Yes, normally the dates 8/01 – 8/21 see a drop in solar declination as the sun moves south. We needed to check that this caused very small spread. The essence of most of these 'non-result' plots is that our assumptions are OK.*

References:
Hawkins, Ed. "Graphics: Scrap rainbow colour scales." Nature 519, no. 7543 (2015): 291.

Anonymous Referee #2
Received and published: 9 April 2018

The manuscript presents the development of a new protocol for evaluating photochemistry and reactivity of air masses comparing different models with in situ measurements of atmospheric composition. The authors use synthetic in situ measurements to demonstrate the feasibility and application of their protocol. The topic of the manuscript is very timely given the wide availability of atmospheric chemistry models and observations and the growing interest in such models and observations for air quality and climate applications. I recommend publication subject to the following comments:

General comments:

The manuscript is sometimes very heavy in terms of the text and some adjustments to better illustrate the points would be beneficial. Some of the figures in the supplement may be better suited in the main manuscript, particularly in Section 2 describing the overall comparison of the models in terms of the key species and the profiles of their reactivity. I suggest moving some of these to the main manuscript.
*Yes. Some of the text adjustments have been made in our re-reading and also in response to RC1's specific comments. We have moved all of the critical model-model comparisons to the main text (the covariance Table S2, Figures S2, S6, S8) and left the sensitivity analysis (with the mostly, hoped-for, null results) in the Supplement.*

In the summary the authors refer to the ability of 3-D models to separate the effects of photochemistry and emissions, for example, and a brief comment on models that can do this, and the advantage of their protocol over such models could be beneficial to the reader.
*We are not sure quite what is wanted here, but that topic (at the end of paragraph #2 in the Summary) has been split out, better explained.*

It is clear the use of this protocol for identifying and evaluating inter-model differences is very useful and some bullet points on the main highlights of the findings in this regard could be helpful.
*Hmm. We have added a short paragraph at the end: "This new protocol for multi-model evaluations helps identify and provide insights into inter-model differences, as well as providing a direct link with measurements made at a much finer scale than the model.*

Specific comments:
Page 8, line 218: 'x' should be subscript in HOx
*done, thanks.*

Page 13, lines 303-304: should "and find out" be "to find out"?
*Yes, this is better- done.*

Page 13, line 309: the beginning of the sentence "For most all models" doesn't make a lot of sense, please clarify

*Yes, identified by both reviewers – this has been changed to "For most models, these reactivity curves…*

Page 15, line 367: should "notable" be "notably"?

*Thanks, done.*

Page 18, line 412: it isn't clear what is meant in describing the selected day – does 8/16 refer to a generic or specific day in August 2016? This is also referred to later in the manuscript.

*Our original notation did not work well when included as text. RC1 also had trouble with this. We have explicitly stated the date in the UCI model and dropped references to day 4 or 'd4' (which happened to be the 4$^{th}$ of the 5-day sequence August 1, 6, 11, 16, 21 that were sued to assess cloud variability. This should be throughout the ms.*

**How well can global chemistry models calculate ~~Chemical modeling of~~ the reactivity of short-lived greenhouse gases in the remote troposphere, knowing the chemical composition~~: a model inter-comparison prescribing a well-measured, remote troposphere~~**

**>>>>please note that the figures have been updated per reviewers to have better clarity, units on the axes, and consistent colors across the models. These changes are not 'tracked', just the updated figure is shown<<<<<<**

Michael J. Prather[1], Clare M. Flynn[1], Xin Zhu[1], Stephen D. Steenrod[2,3], Sarah A. Strode[2,3], Arlene M. Fiore[4], Gustavo Correa[4], Lee T. Murray[5], Jean-Francois Lamarque[6]

[1]Department of Earth System Science, University of California, Irvine, CA 92697-3100, USA

[2]NASA Goddard Space Flight Center, Greenbelt, MD, USA

[3]Universities Space Research Association (USRA), GESTAR, Columbia, MD, USA

[4]Department of Earth and Environmental Sciences and Lamont-Doherty Earth Observatory of Columbia University, Palisades, NY, USA

[5]Department of Earth and Environmental Sciences, University of Rochester, Rochester, NY 14627-0221, USA

[6]Atmospheric Chemistry, Observations and Modeling Laboratory, National Center for Atmospheric Research, Boulder, CO 80301, USA

Correspondence to: Michael J. Prather (mprather@uci.edu)

*Abstract.* We develop a new protocol for merging in situ measurements with 3-D model simulations of atmospheric chemistry with the goal of integrating over these data to identify the most reactive air parcels in terms of tropospheric production and loss of the greenhouse gases ozone and methane. Presupposing that we can accurately measure atmospheric composition, we examine whether models constrained by such measurements agree on the chemical budgets for ozone and methane. In applying our technique to a synthetic data stream of 14,880 parcels along 180W, we are able to isolate the performance of the photochemical modules operating within their global chemistry-climate and chemistry-transport models, removing the effects of modules
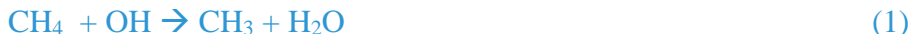
1

controlling tracer transport, emissions, and scavenging. Differences in reactivity across models are driven only by the chemical mechanism and the diurnal cycle of photolysis rates, which are driven in turn by temperature, water vapor, solar zenith angle, clouds, and possibly aerosols and overhead ozone, which are calculated in each model. We evaluate six global models and identify their differences and similarities in simulating the chemistry through a range of innovative diagnostics. All models agree that the more highly reactive parcels dominate the chemistry (e.g., the hottest 10% of parcels control 25-30% of the total reactivities), but do not fully agree on which parcels comprise the top 10%. Distinct differences in specific features occur, including the spatial regions of maximum ozone production and methane loss, as well as in the relationship between photolysis and these reactivities. Unique, possibly aberrant, features are identified for each model, providing a benchmark for photochemical module development. Among the 6 models tested here, 3 are almost indistinguishable based on the inherent variability caused by clouds, and thus we identify 4, effectively distinct, chemical models. Based on this work, we suggest that water vapor differences in model simulations of past and future atmospheres may be a cause of the different evolution of tropospheric $O_3$ and $CH_4$, and lead to different chemistry-climate feedbacks across the models.

## 1. Introduction

The daily passage of sunlight through the lower atmosphere drives photochemical reactions that control many short-lived greenhouse gases (GHGs) and other pollutants. This daily cycle occurs across a range of different chemical compositions; such that even neighboring air parcels can exhibit a wide range in their reactivity with respect to GHGs (Prather et al., 2017; henceforth P2017). This paper selects a tomographic sampling of air parcels from a high-resolution chemistry-transport model, meant to simulate what an aircraft mission might measure (e.g., NASA's Atmospheric Tomography Mission: ATom, 2017), and asks if a cohort of six global chemistry models can agree on the reactivity of these parcels. To do this, we develop a new protocol and set of diagnostics for merging in situ measurements with 3-D model simulations of atmospheric chemistry. We focus here on tropospheric ozone production and loss (P-O3, L-O3) and methane loss (L-CH4), as these two gases are the most important GHGs controlled through tropospheric chemistry. Further, control of $CH_4$ and $O_3$ provides an important pathway for

2

limiting near-term climate change (Shindell et al., 2012). These reactivities are defined in terms of specific rates:

$$CH_4 + OH \rightarrow CH_3 + H_2O \qquad\qquad (1)$$
$$HO_2 + NO \rightarrow NO_2 + OH \qquad\qquad (2a)$$
$$RO_2 + NO \rightarrow NO_2 + RO \qquad\qquad (2b)$$
$$O_3 + OH \rightarrow O_2 + HO_2 \qquad\qquad (3a)$$
$$O_3 + HO_2 \rightarrow HO + O_2 + O_2 \qquad\qquad (3b)$$
$$O(^1D) + H_2O \rightarrow OH + OH \qquad\qquad (3c)$$

L-O3 is rate (1); P-O3 is rates (2a) + (2b); L-O3 is rates (3a) + (3b) + (3c). All of the analysis here occurs at pressures > 200 hPa and thus the P-O3 term from photolysis of $O_2$, important at 100-200 hPa in the tropics, can be ignored. A definition and example of these reactivities, How reactivities and how they can be calculated for assigned to an air parcel, is found in P2017 and the Supporting Information to this paper.

From the early model-and-measurement assessments that were initiated to support the stratospheric ozone assessments (NAP, 1983; NASA, 1993), through to the most recent multi-model evaluations of atmospheric chemistry to be used in upcoming climate assessments (Collins et al., 2017; Morgenstern et al., 2017; Eyring et al., 2006; Myhre et al., 2017), there is one truism: the models always produce different results even when they agree upon the protocols, and intend to do the same simulation. For assessments one seeks common ground to find a robust result; whereas for science one seeks a cause of disagreement to identify how models can be improved. This paper focuses on the latter. Given the scale and complexity of current 3D global chemistry models, potential causes of differences in model-simulated distributions of chemical tracers are many is large. The numerical algorithms and parameterizations for the transport, mixing, and thus dispersion of emissions is clearly one cause (Prather et al., 2008; Lauritzen et al., 2014; Orbe et al., 2016); while photochemical mechanisms that produce and destroy species are another (Olson et al., 1997; PhotoComp, 2010).

This paper initiates a new technique for multi-model comparison that uses prescribed initial chemical composition of air parcels, which we refer to as the modeling data stream. We presuppose that we can accurately measure or otherwise know atmospheric composition, and then ask if models calculate the same global chemical budgets for ozone and methane. Our

approach eliminates many of the factors that drive model differences and allows us to focus on the photochemical reactivities as integrated over a day. Instantaneous reactivities can be inferred from measurements of reactive chemical species and the radiation field combined with

100    laboratory cross sections and reaction rate coefficients, e.g., (Olson et al., 2012). Attempts to follow the chemical evolution of air parcels with aircraft measurements is limited and quasi-Lagrangian at best (Nault et al., 2016). Even the concept of isolated Lagrangian parcels is limited, since parcels shear and mix rapidly as they go from a large, chemically coherent air mass to a heterogeneous mix of smaller features (Batchelor, 1952; Prather and Jaffe, 1990). Yet,

105    simulating the photochemical changes in $CH_4$ and $O_3$ requires integration over the daily cycle of photolytic rates, which change greatly and irregularly over the day based on the interaction of the sun and cloud systems. Unfortunately, there is no known approach to track and measure the 24-hour net change in ozone or methane for an air parcel in the free troposphere. Here and in P2017, we approximate the reactivity of an air parcel by running our global chemistry models

110    with their regular meteorology and chemical modules, but with transport and mixing of tracers shut down to keep the grid cells isolated. Effectively, we are able to use the standard full 3D model as a collection of box models (i.e., one per grid cell), while incorporating its diurnal cycle of photolysis and cloud fields. Such simulations, named the A-runs, are artificial since real air parcels constantly move and mix with their environment. Statistical comparison of A-run

115    reactivities from the six models with those using the standard 3D versions is examined in P2017, and shows agreement with some minor biases due to the A-run formulation.


The participating models and the modeling data stream are described in Section 2. This effort was completed before the release of the ATom aircraft data (ATom, 2017) and thus we use a

120    1/2°-resolution model to generate the data stream. Section 3 presents and compares the statistics of P-O3, L-O3, and L-CH4 and J-values from the 14,880 parcels, including 5 different days in August to sample variability in cloud systems. Sorted distributions show the models' agreement on the most highly reactive parcels. The final discussion in Section 4 considers the role of inherent uncertainty in modeling parcel reactivity, of basic differences in the models, and

125    whether the new statistics developed here identify and characterize differences in the photochemical modules. For insight on the most reactive air parcels of the remote troposphere, we await a repeat of this work with the ATom data stream.

## 2. Chemistry models and simulations.

130

The six global chemistry models here are basically the same as those in P2017: Geophysical Fluid Dynamics Laboratory (GFDL), Goddard Institute for Space Studies (GISS), Goddard Space Flight Center (GSFC), GEOS-Chem (GC), National Center for Atmospheric Research (NCAR), and UC Irvine (UCI). For model versions and updates, see Tables 1 and S1ab.

135

A model-simulated data stream of air parcels was prepared from an older version of the UCI model (v72a) with higher than usual resolution (T319L60, ~0.55 degrees) and sampled at 00UT 15 August 2005 at aircraft flight levels along 3 meridians next to 180E. All the model grid cells are used with no attempt to follow ATom profiling. This set of 14880 points is similar in

140 number to 10-second data from an aircraft mission logging 50 flight hours in the Pacific basin, such as each seasonal deployment of ATom. Prescribed species are: $O_3$, NOx (=NO+$NO_2$), $HNO_3$, $HNO_4$, PAN (peroxyacetyl nitrate), $RNO_3$ ($CH_3NO_3$ and all alkyl nitrates), HOOH, ROOH ($CH_3OOH$ and smaller contribution from $C_2H_5OOH$), HCHO, $CH_3CHO$ (acetaldehyde), $C_3H_6O$ (acetone), CO, $CH_4$, $C_2H_6$, alkanes (all $C_3H_8$ and higher), alkenes (all $C_2H_4$ and higher),

145 aromatics (benzene, toluene, xylene), $C_5H_8$ (isoprene plus terpenes), plus temperature (T) and specific humidity (q). Zonal mean latitude by pressure plots of $O_3$, CO, HCHO, NOx, PAN and q are shown in Figure S1.

The implementation of this data stream of reactive species is model dependent. All models begin
150 with their own 3D initialization data set that is used to restart a model simulation beginning on August 16. The specified air-parcel NOx, for example, will be initialized as separate NO and $NO_2$ abundances by scaling the model's restart values for NO and $NO_2$ to match the specified parcel NOx. Similarly, a single value for aromatics will be partitioned over benzene, toluene, and xylene by models that resolve these species in accord with the restart values. The models
155 place each parcel (i.e., overwrite the restart values) in the grid cell containing the latitude, longitude, and pressure specified for that parcel. If that preferred grid cell is already occupied with an air parcel, then an alternate adjacent grid cell is selected. It is recommended that alternate cells be shifted to minimize the change in photolytic environment (e.g., shift by

5

longitude but maintain surface albedo and atmospheric mass).  ~~See Supplement for how these~~

160 ~~data were implemented in the models and how T~~two chemistry-climate models (GFDL, NCAR) were unable to completely overwrite the modeled T&q values with ~~those from the~~ data stream values (see sensitivity tests below).  See also Supplement for additional details.

Implications for reactivities are discussed below.  It is difficult, if not impossible, to specify 24-

165 hour cloud fields, from observations or a model, in a way that all models here could implement consistently.  Treatment of photolysis rates in uniform cloud layers is still quite different across models, and fractional overlapping cloud fields are often ignored, e.g., (Prather, 2015). Likewise, we do not attempt to control the profiles of $O_3$ and aerosol above and below the air parcels insofar as they impact photolysis.  Hence we diagnose photolysis rates (J values) in

170 addition to reactivities.

An inherent uncertainty is the day-to-day variability of clouds experienced by each parcel.  Thus for the single data stream, each model calculates reactivities using the same chemical initialization but beginning with 5 different days in August: $1^{st}$, $6^{th}$, $11^{th}$, $16^{th}$, and $21^{st}$.  This 5-

175 day variance gives us a measure of the uncertainty due to cloud variability, is similar across models, and thus provides a lower limit on the detection of model-model differences, i.e., a measure of as-good-as-it-gets in this comparison.

Several uncertainties are not answered with the standard protocol of 5-day runs:  Models ran

180 with different calendar years and so how do 5-day means vary from year to year? Does the changing solar declination matter?  Will different restart files (affecting $O_3$ and aerosol profiles) alter the results? What if the 24h integrations began at midnight rather than noon?  How different are the CCMs because they use their own T&q for the parcels?  The UCI CTM ran additional sensitivity calculations to address these questions, see Section 3.5 and Supplementary Figures.

185

| **Table 1.** Participating models | | | | | | |
|---|---|---|---|---|---|---|
| | model | type | meteorology | T & q | POC | model grid |
| GFDL | AM3 | CCM | NCEP (nudged) | CCM | Arlene Fiore | C180 x L48 |
| GISS | GISS-E2.1 | CCM | daily SSTs, nudged to MERRA | parcel | Lee Murray | 2º x 2.5º x 40L |
| GSFC | GMI-CTM | CTM | MERRA | parcel | Sarah Strode | 1º x 1.25º x 72L |
| GC | GEOS-Chem | CTM | MERRA-2 | parcel | Lee Murray | 2º x 2.5º x 72L |

| NCAR | CAM4-Chem | CCM | MERRA | CCM | Jean-Francois Lamarque | 0.47°x0.625°x52L |
| UCI | UCI-CTM | CTM | ECMWF IFS Cy38r1 | parcel | Michael Prather | T159N80 x L60 |

### 3. Reactivity across the models

190     The difference in modeled reactivities for each parcel combines variations in cloud fields with basic differences in the chemical models (i.e., chemical mechanisms, numerical methods, photolysis treatment of cloudy and clear sky). The 5-day means reduce the effect of cloud variations but leave the fundamental differences in the photochemical modules, both photolytic and kinetic reactions. Our comparison looks at the parcel-by-parcel differences including the

195     scatter (root mean square (rms) rms differences) and average values across the models. To provide a standard for comparisons, we seek a reference case based on several models, and this is easily identified with the rms differences across all model pairs in (Table S2). UCI ran 3 different model years to determine estimate the a lower limit rms value caused by interannual variability (blue in Table 2), i.e., when the cross-model differences approach this limitvalue, we

200     can accept that the photochemical modules including clouds cannot be said to be different in this study. For the reactivities (P-O3, L-O3, L-CH4), none of the cross-model pairs reached this lower limit, but certain groupings were consistently close, within a factor of 2 of this limit. For L-O3 and L-CH4, any pair of GSFC-GC-UCI fall within this range, while GFDL, GISS and NCAR are a factor of 5-10 above it. For the two CCMs this is likely caused by their use of

205     different T&q's, while for GISS it probably lies in the chemical model. For P-O3, only the pair GC-UCI is within a factor of 2, but GFDL-GSFC-GC-UCI form a distinct cluster. The J values, J-O1D ($O_3$+hv=>$O_2$+O($^1$D)) and J-NO2 ($NO_2$+hv=>NO+O), show groupings similar to this cluster, reflecting their common use of Fast-J versions (Wild et al., 2000; Prather, 2015), although this is unlikely to explain their similarity in P-O3.

210

Based on the average of the 5-day parcel means, we find a cluster of 3 similar models and 3 independent models. We need to find a common reference case against which to plot and statistically evaluate the models. Rather than pick one model, we take the 3-model average, GSFC-GC-UCI, as our reference. This clustering may be due to similar heritage: GSFC and GC

215     are derived from a common tropospheric chemistry module; all 3 models and GISS have a

7

common heritage for photolysis module.  In the comparisons below, we will use terms like 'bias' to describe differences with respect to this reference model.  Such biases are not meant to be model errors since we do not know the correct answer; they are just model-model differences.

**Table 2.** RMS differences of 5d-mean parcels across model pairs

| model /model | GFDL | GISS | GSFC | GC | NCAR | UCI | U2015 | U1997 |
|---|---|---|---|---|---|---|---|---|
| **P-O3 = 0.79 ppb/d** | | | | | | | | |
| GFDL | 0 | 0.87 | **0.19** | **0.15** | **0.15** | **0.16** | *0.15* | *0.16* |
| GISS | | 0 | 0.90 | 0.83 | 0.81 | 0.80 | *0.80* | *0.79* |
| GSFC | | | 0 | **0.17** | 0.24 | **0.19** | *0.20* | *0.20* |
| GC | | | | 0 | **0.17** | **0.10** | *0.11* | *0.11* |
| NCAR | | | | | 0 | **0.15** | *0.14* | *0.14* |
| UCI | | | | | | 0 | ***0.06*** | ***0.06*** |
| U2015 | | | | | | | 0 | ***0.06*** |
| U1997 | | | | | | | | *0* |
| **L-O3 = 1.45 ppb/d** | | | | | | | | |
| GFDL | 0 | 1.86 | 0.88 | 0.86 | 0.92 | 0.96 | *0.97* | *0.98* |
| GISS | | 0 | 1.06 | 1.06 | 1.26 | 1.02 | *1.01* | *1.02* |
| GSFC | | | 0 | **0.20** | 0.67 | **0.21** | *0.24* | *0.25* |
| GC | | | | 0 | 0.59 | **0.22** | *0.24* | *0.26* |
| NCAR | | | | | 0 | 0.68 | *0.71* | *0.71* |
| UCI | | | | | | 0 | ***0.12*** | ***0.12*** |
| U2015 | | | | | | | 0 | ***0.13*** |
| U1997 | | | | | | | | *0* |
| **L-CH4 = 0.63 ppb/d** | | | | | | | | |
| GFDL | 0 | 0.38 | 0.23 | 0.22 | 0.20 | 0.26 | *0.27* | *0.27* |
| GISS | | 0 | 0.42 | 0.44 | 0.47 | 0.44 | *0.45* | *0.45* |
| GSFC | | | 0 | **0.09** | 0.25 | **0.10** | *0.12* | *0.12* |
| GC | | | | 0 | 0.22 | **0.09** | *0.10* | *0.11* |
| NCAR | | | | | 0 | 0.25 | *0.26* | *0.26* |
| UCI | | | | | | 0 | ***0.06*** | ***0.06*** |
| U2015 | | | | | | | 0 | ***0.06*** |
| U1997 | | | | | | | | *0* |
| **J-NO2 = 4.45 (e-3 /s)** | | | | | | | | |
| GFDL | 0 | 1.14 | **0.43** | **0.42** | 0.58 | 0.71 | *0.72* | *0.74* |
| GISS | | 0 | 1.11 | 1.02 | 0.96 | 0.78 | *0.82* | *0.80* |
| GSFC | | | 0 | **0.37** | 0.55 | **0.65** | *0.71* | *0.72* |
| GC | | | | 0 | **0.51** | **0.56** | *0.60* | *0.63* |
| NCAR | | | | | 0 | **0.62** | *0.62* | *0.65* |
| UCI | | | | | | 0 | ***0.33*** | ***0.33*** |
| U2015 | | | | | | | 0 | ***0.34*** |
| U1997 | | | | | | | | *0* |
| **J-O1D = 1.19 (e-5 /s)** | | | | | | | | |
| GFDL | 0 | 1.07 | **0.15** | **0.15** | 0.25 | **0.17** | *0.17* | *0.18* |
| GISS | | 0 | 1.06 | 1.09 | 0.94 | 1.02 | *1.02* | *1.01* |
| GSFC | | | 0 | **0.12** | 0.24 | **0.13** | *0.15* | *0.16* |
| GC | | | | 0 | 0.29 | **0.17** | *0.17* | *0.18* |
| NCAR | | | | | 0 | 0.24 | *0.24* | *0.24* |
| UCI | | | | | | 0 | ***0.08*** | ***0.08*** |
| U2015 | | | | | | | 0 | ***0.08*** |

| | GFDL | GISS | GSFC | GC | NCAR | UCI | U2015 | U1997 |
|---|---|---|---|---|---|---|---|---|
| U1997 | | | | | | | | *0* |

### 3.1. Average profiles

Altitude profiles of reactivities and J values averaged over 24 hours, 5 days in August, and latitude blocks (50S-20S, 20S-20N and 20N-50N) are shown in Figure 1S2 (6 models, 3 blocks, 18 profiles per panel). As expected for August, the 50S-20S values are very low, while the 20S-20N and 20N-50N ones are equally high. This basic latitude-season pattern holds across all models. The variability across the 5 separate days in the UCI model (Figure S2S3) is primarily a smooth trend through August reflecting the changing solar declination from 18° to 12°, but instances of highly variable cloud fields occur, even when averaged over 30° in latitude.

For J-O1D, five models (GFDL, GSFC, GC, NCAR, UCI) agree well over all pressures and latitude blocks, but NCAR is, unusually, 10% higher only in the 20S-20N block. J-O1D from GISS is 80% larger than other models for all pressure and latitude blocks, but this does not translate directly or simply into reactivities, where GISS L-O3 is higher (expected) but L-CH4 is lower (unexpected). For J-NO2, model differences are not so great and show largest values at 20N-50N consistent with the longer summer daytime hours. The spread in J-NO2 is partly understandable because of ambiguous choices in interpolating the temperature dependence of recommended $NO_2$ cross sections and quantum yields (i.e., the absorption cross sections are given at 220K and 294K; the quantum yields, at 248K and 298K; and the choice of whether to interpolate linearly or logarithmically, or whether to extrapolate or not, affects J-NO2, especially in the upper troposphere). This ambiguity does not exist for J-O1D recommended cross section and quantum yields. J-O1D is strongly dependent on the overhead $O_3$ column, and the zonal mean total $O_3$ column from the models is compared with recent satellite measurements in Figure S3S4. NCAR's $O_3$ column is anomalously lower only in the 20S-20N region and likely explains their higher J-O1D noted above.
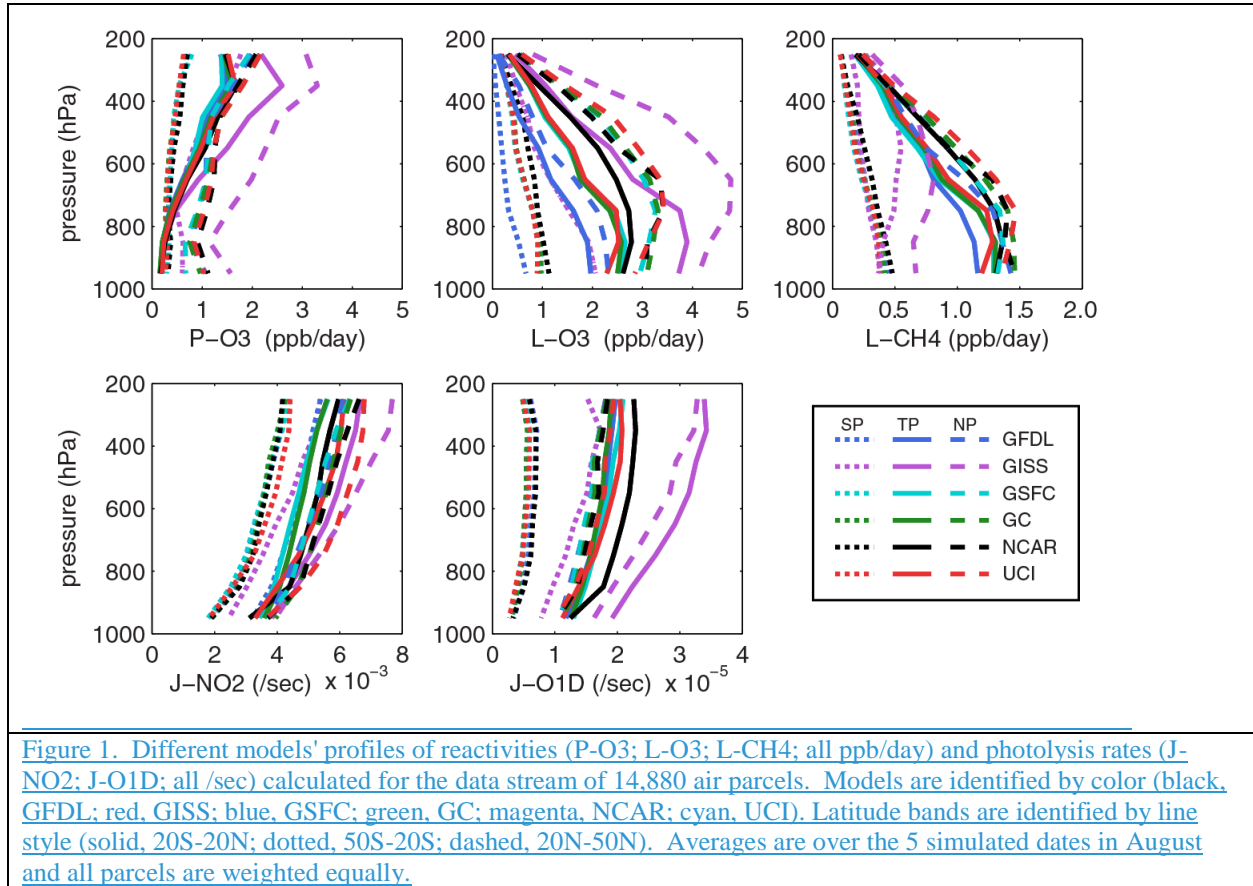
Reactivity profiles for the 5 non-GISS models show excellent agreement for P-O3 but noticeable differences for L-CH4 and even larger ones for L-O3 (Figure 1S2). The altitude profiles are similar for the 5 models, indicating that the cause of the L-O3 spread is likely related to $HO_X$.

250  The GISS results are anomalous, with much higher P-O3 and an L-O3 vs. L-CH4 relationship

that seems counter to known chemistry in which both L-O3 and L-CH4 maximize with the high

HOxx values in the warmer, wetter, lower troposphere of the tropical Pacific.



Figure 1. Different models' profiles of reactivities (P-O3; L-O3; L-CH4; all ppb/day) and photolysis rates (J-NO2; J-O1D; all /sec) calculated for the data stream of 14,880 air parcels. Models are identified by color (black, GFDL; red, GISS; blue, GSFC; green, GC; magenta, NCAR; cyan, UCI). Latitude bands are identified by line style (solid, 20S-20N; dotted, 50S-20S; dashed, 20N-50N). Averages are over the 5 simulated dates in August and all parcels are weighted equally.

255  *3.2. 14,880 parcels*

We examine the relationship between the 3 reactivities in each model with scatter plots of P-O3

and L-CH4 against L-O3 in Figure 21. Each plot has 14,880 points (5-day parcel means) and is

split by location: 60S-20S & 20N-60N (extra tropics, gray); tropics upper (20S-20N, p < 600

260  hPa, cyan) and lower (p > 600 hPa, blue). Percentiles ($10^{th}$, $50^{th}$, $90^{th}$) in each dimension are

plotted as red dash-dot lines, and thus most points in the well correlated L-CH4 vs. L-O3 lie

along the 3 quasi-diagonal intersections of red lines. The right-angle separation of high P-O3

and high L-O3 in the tropics reflects the high NOx (P-O3) in this data stream is in the upper

troposphere and the largest L-O3 is from wet environments of the lower troposphere. GFDL has

265    the most compact distribution of parcels and GISS, the most scattered.  Four models (GSFC, GC, NCAR, UCI) have remarkably similar patterns in terms of the percentiles and structure, e.g., for L-CH4 vs. L-O3 they show the lower tropics dominating the upper part of the distribution and the extra-tropics, the lowermost points.  GFDL has similar percentiles for P-O3 and L-CH4, but a much smaller spread for L-O3 that explains their compacted scatter plots.  GISS is unique with

270    much larger spread in both P-O3 and L-O3 but a compressed distribution in L-CH4.  From these scatter plots, we can say that the 4 models are remarkably consistent, that GFDL is similar but should reexamine their L-O3 diagnostic, and that GISS has a 'uniqueness' in its L-O3 vs. L-CH4 relationship as well as large scatter in both P-O3 and L-O3.  While consistency does not guarantee correct implementation of the photochemical model (i.e., rate coefficients, cross

275    sections), uniqueness is something that needs more investigation as it may be an error or may lead to fixes in the 'consistent' models.  Scatter plots of J-NO2 and J-O1D vs. L-O3 (Figure S4S5) show similar J-value statistics for the 5 non-GISS models, and all models show a similar location of the 3 sets of points (extra-tropics, lower-tropics, upper-tropics) within their own percentiles.

280

On a parcel-by-parcel basis we compare in Figure 32 the 5-day means from all 6 models against the reference case for the 3 reactivities and 2 J-values.  If the models were all alike, they would fall tightly on the 1:1 line (black dashed).  In each panel there are 89,280 points with many overlapping.  The order of plotting (shown by the legend) is important for visual impression

285    since the latter points often overlie the earlier ones and the choice of order was based partly on the rms differences, with greatest first and smallest last.  Here we can clearly see the type of scatter, the pattern of discrepancies across models, and at what levels of reactivity such discrepancy it occurs.  It provides a focus for model development:  UCI should reexamine its J-NO2 at the higher values and its P-O3 in the 1-3 ppb/day range; NCAR should examine why it

290    has so much scatter in L-O3 and L-CH4 (see discussion of T&q later); GFDL has similar scatter (see T&q) but also has a low-bias in L-O3; and GISS has many differences that can be examined.  As a cross-model question, are the above-the-line (UCI) and below-the-line (GSFC) differences in P-O3 and L-CH4 related to the same pattern in J-NO2?

295 A simple summary of these statistics – averages and rms differences relative to the reference case – is given in Table 32. We have selected (bold italics) those entries that seem anomalous as also found in Figure 23. For example, average P-O3 ranges from 0.77 to 0.84 ppb/day for 5 models but is 1.40 ppb/day for GISS. Likewise, average L-O3 ranges from 1.44 to 1.54 ppb/day for 4 models, but is 0.83 for GFDL and 2.25 ppb/day for GISS. The rms differences with respect

300 to the reference case favors the 3 models that define that case, but also shows that GFDL and NCAR are close to the reference case for P-O3, but farther away for L-O3 and L-CH4 probably caused by their T&q values (see later).

**Table 32.** Average Reactivities and Standard Deviations w.r.t. reference case (average of 3 #'d models).

| Reactivity | P-O3 (ppb/d) | L-O3 (ppb/d) | L-CH4 (ppb/d) | J-NO2 (e-3 /s) | J-O1D (e-5 /s) |
|---|---|---|---|---|---|
| *a) Average Reactivities (5-day averages of 14,880 parcels)* | | | | | |
| Reference Case# | 0.792 | 1.449 | 0.633 | 4.452 | 1.194 |
| GFDL | 0.771 | *0.826* | *0.579* | 4.237 | 1.177 |
| GISS | *1.405* | *2.248* | *0.429* | *5.159* | *2.154* |
| GSFC# | 0.755 | 1.436 | 0.610 | 4.258 | 1.194 |
| GC# | 0.793 | 1.444 | 0.641 | 4.392 | 1.164 |
| NCAR | 0.839 | 1.541 | 0.666 | 4.475 | 1.305 |
| UCI# | 0.827 | 1.467 | 0.648 | *4.705* | 1.224 |
| *UCI 2015* | *0.833* | *1.474* | *0.651* | *4.725* | *1.227* |
| *UCI 1997* | *0.833* | *1.471* | *0.649* | *4.724* | *1.231* |
| *b) RMS Differences versus Reference Case, using 5-day means* | | | | | |
| GFDL | 0.14 | *0.89* | *0.23* | 0.44 | 0.13 |
| GISS | *0.84* | *1.04* | *0.43* | *0.93* | *1.05* |
| GSFC# | 0.12 | 0.11 | 0.05 | 0.30 | 0.06 |
| GC# | 0.07 | 0.12 | 0.05 | 0.23 | 0.08 |
| NCAR | 0.16 | *0.64* | *0.23* | 0.47 | *0.24* |
| UCI# | 0.08 | 0.13 | 0.06 | 0.39 | 0.09 |
| *UCI 2015* | *0.10* | *0.17* | *0.08* | *0.48* | *0.11* |
| *UCI 1997* | *0.10* | *0.18* | *0.08* | *0.50* | *0.12* |
| *c) RMS Differences day-to-day versus 5-day mean of same model* | | | | | |
| GFDL | 0.09 | 0.26 | 0.12 | 0.36 | 0.08 |
| GISS | *0.53* | *0.41* | 0.08 | 0.67 | *0.29* |
| GSFC | 0.13 | 0.22 | 0.10 | 0.48 | 0.12 |
| GC | 0.09 | 0.19 | 0.08 | 0.43 | 0.10 |
| NCAR | *0.15* | *0.54* | *0.21* | 0.62 | 0.18 |
| UCI | 0.09 | 0.18 | 0.08 | 0.52 | 0.12 |
| *UCI year-to-year* | *0.06* | *0.12* | *0.06* | *0.33* | *0.08* |
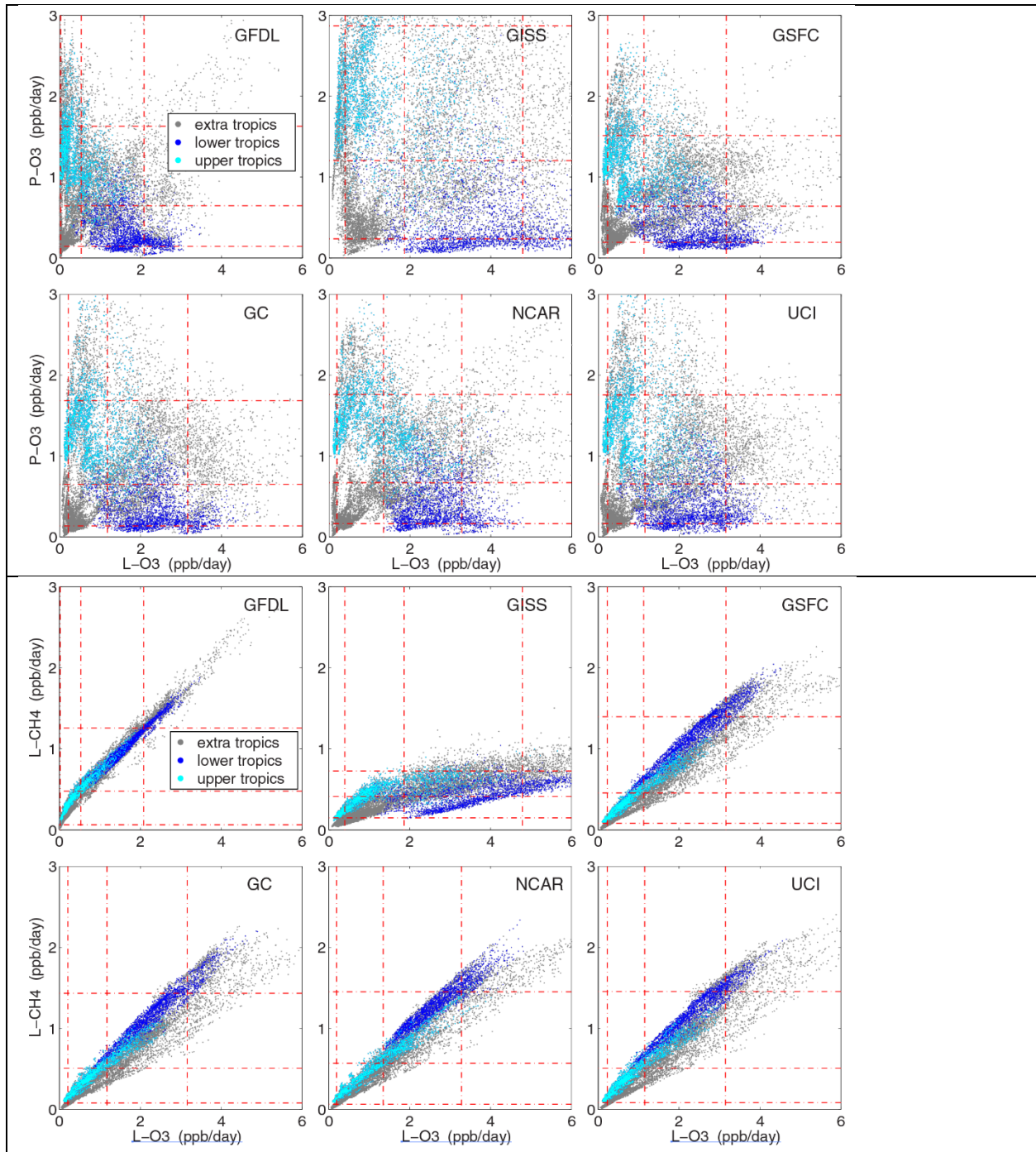
Figure 2̶1̶. Parcel reactivities of (top) P-O3 and (bottom) L-CH4 vs. L-O3 for each of the models. Points are colored by location: 60S-20S & 20N-60N (extra tropics, gray); tropics (20S-20N) upper (p < 600 hPa, cyan) and lower (p > 600 hPa, blue). the 10th, 50th, and 90th percentiles in each dimension are plotted as red dash-dot lines.
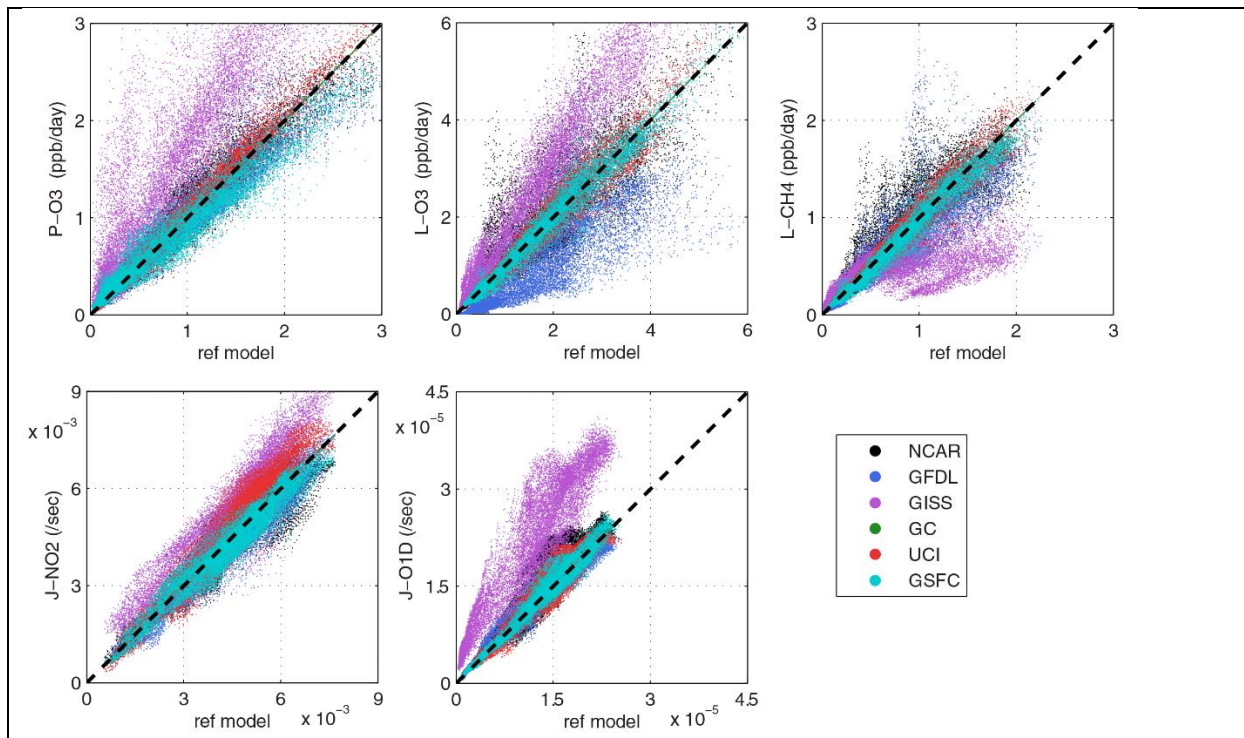
Figure 32. Direct parcel-by-parcel comparison of modeled reactivities (a, P-O3; b, L-O3; c, L-CH4; all ppb/day) and photolysis rates (d, J-NO2; e, J-O1D; all /sec) calculated for the 14,880 simulated air parcels. Each point is an average over the 5 simulated dates in August (8/01, 8/06, 8/11, 8/16, 8/21). The 1:1 line is shown (black dashed) for each plot. The reference values (X axis) are the average of 3 similar models (GSFC, GC, UCI) selected by examining the rms differences across all the models (see text). ~~Note that the model points are plotted successively on top of one another and thus the earlier plotted models may appear less frequent: in order, NCAR (black), GFDL (magenta), GISS (cyan), GC (blue), UCI (red), GSFC (green).~~ For this plot alone, m~~M~~odels are plotted in the following order with the most disperse points being first for visibility: NCAR, GFDL, GISS, GC, UCI. GSFC. The model colors throughout this paper are consistent, but the order of plotting is shown in the legend.

310

### 3.3. 5 days versus 5-day mean

The variability of the 5 days in August tells us about the synoptic variability of clouds and possibly O₃ columns in each model. The rms difference between the 5 individual days and the 5-

315    day mean for each model (Table 3c2) shows ~~that~~ that GISS and NCAR have much larger variability in reactivities, caused by and mirrored by those in J values. These rms differences in J values for GISS and NCAR are surprising. Collectively, we should reexamine this variability in all the models to ascertain its cause. In general, the slopes of the individual versus reference model for reactivities are close to 1 (Table S23) because the slope is determined by the large

320    gradients with latitude and pressure that most models agree on. In comparing individual days versus 5-day mean, it is encouraging that this slope averages 1±0.04 for all reactivities and

14

models (using each model's 5-day mean as its reference case, Table S~~3~~4). Also, the slope decreases from about 1.01 to 0.96 through August as expected with declining photolysis rates in the north.

325

The rms difference across the 5 days is also a measure of how well the 5-day parcel mean can represent the true chemical model. Assuming that the cloud variability is random, the 5-day means with respect to other models are not really different unless that model-model rms exceeds some fraction of the day-to-day rms of the models involved. Using the UCI test with different

330    model years, we find that the year-to-year rms differences are about 2/3 of the day-to-day rms over 5-days. Thus, we cannot be sure that the rms differences between NCAR and ~~model versus the~~ reference case ~~rms values for NCAR~~ are due to the inadequacy of the 5-day mean to represent the mean NCAR chemistry model (Tables 3b~~2a&~~c). On the other hand, some other source of model error is likely responsible for the large day-to-day rms.

335

*3.4 The "hot" air parcels*

Following the "which air matters" theme of P2017, we look at the more reactive air parcels ~~and~~

340    to find out if the models agree on these. For each reactivity, we sort the 5-day parcel means in increasing order and integrate the cumulative reactivity. The value at 100% (all 14,880 parcels) is equal to the average reactivity of the sample (Table 3~~2~~a), and this is renormalized to 1 for comparison across models (Figure 4~~S6~~, Table S4). With sorting, these curves must be monotonic and convex. The steeper the curve, the more important the top reactive parcels are in

345    determining the total. For most ~~all~~ models, these ~~and~~ reactivity~~ies these~~ curves are remarkably similar and fall within the range seen for 5 different days ~~in~~ with the same model (UCI, Figure S5~~7~~, Table S5). Focusing on the upper 10%, the outliers are unusual and reactivity specific: ~~for P-O3, one group (GISS-GSFC-NCAR) are less steep than the other (GFDL-GC-UCI), and this splits the reference-case models;~~ for L-O3, GFDL is much steeper that the other models,

350    consistent with the feature identified earlier in the scatter plots; and for L-CH4, GISS is much shallower. Surprisingly, with ~~less steep although NCAR also deviates slightly in this way from~~

15

the others.  In this diagnostic , GISS is not an obvious such a clear outlier for P-O3 and L-O3 as seen in previous comparisons.

355 From this cumulative reactivity figure, one can see that the top 5% of parcels comprise 15% of the total reactivity, effectively a slope of 3:1.  With the exceptions noted, total reactivity for the top 5-10-25-50% of the parcels (Table S45) is similar across models and across days within a model (Table S56).  Focusing on the top 10% of parcels for each reactivity, we plot their latitude-by-pressure distribution for each model in Figure 53.  Top P-O3 are in the upper

360 troposphere where NOx was highest in the specified data stream; and top L-O3 and L-CH4 are in the lower troposphere associated with warmer temperatures and higher water vapor, with L-CH4 being at lower altitude than L-O3 (all models except GISS).  There is a region of top P-O3 parcels about 40N that extends into the lower troposphere, although the shape varies across models.  The vertical pattern of top-10% parcels about 22°S clearly varies across models with

365 GISS-GC-UCI not selecting these parcels.

Overlap of these three sets of parcels are quantified as Venn diagrams for each model in Figure 64.  Very few top-10 parcels are in the triple-overlap area (1-10%); but when P-O3 parcels coincide with either L-O3 or L-CH4 parcels, they generally lie in this triple-overlap area. The

370 only major exception to this pattern is GISS.  In terms of L-O3 and L-CH4 overlap, 4 models are very consistent (76-80%); but GISS is unusually low (49%) and GFDL is unusually high (93%).  These patterns help identify distinctly different chemistries in these models that have been identified with other diagnostics.  The Venn overlap diagrams will become more interesting with an observational data stream as they point to the co-occurrence of unusual atmospheric parcels.

375

At what level do the models agree on the hot, top-10% parcels?  We use the reference case defined above and sort each reactivity to identify the top-10%, retain those parcel numbers and compare across models.  Table S67 gives each model's overlap of their top-X% parcels in terms of the percent that also occur in the top-X% reference case.  For a range of X, 5%-10%-25%-

380 50%, the overlap increases successively with many models having 90% overlap for the top-50%.  The exceptions are GFDL with lower than typical overlap for L-O3 at all top-X% levels, and GISS, with lower overlap for L-CH4.  This new diagnostic is helpful in understanding these

model differences because it implies that the L-O3 and L-CH4 differences identified previously are not caused by a systematic offset in all parcels, but rather by a selection of different parcels.

385 As expected, the three models GSFC-GC-UCI that define the reference case all have about 90% overlap for the top-10% parcels, and so we do not learn much with this. In terms of linking models with similar chemistries, probably 80% overlap is a good mark, because we see that the different UCI years drop off to 85% in L-O3 and L-CH4. Overlap in P-O3 is much easier to achieve as the few high-NOx parcels drive high P-O3 in all models: at the top-25% parcels, the

390 P-O3 overlap is about 84% or better for all models.

On a day-to-day basis, we examine the top-10% overlap for GC-GSFC-GC-NCAR-UCI models, using their own 5-day mean as the reference (Table S78). Cloud variations across the 5 days lead to overlaps for the top-10% parcels ranging from 78 to 92% at best. NCAR has similar self-

395 overlaps for P-O3 but only 58 to 72% for L-O3 and L-CH4, because the modeled T&q changes with each day in August and greatly reduces the overlap of the hot parcels. This further supports T&q as being important drivers of L-O3 & L-CH4. The use of 5-day calculations with varying cloud fields is essential in identifying the top reactive parcels.

400 We plot the modeled reactivity of individual model 5-day mean parcels in ascending order based on the sorted top-10% parcels in the reference case (Figure 7S8). Hence the reference case (black line) is a monotonically increasing curve; while the individual models produce a scattered distribution of points. As expected, the 3 models defining the reference case have some scatter but mostly overlap with the reference case. UCI is typically higher and GSFC is lower. For J-

405 NO2 in these most reactive parcels, UCI is notablye higher as is GISS, a result seen in the average profiles (Figure 1S2), but it does not affect the reactivities. The mean bias of models relative to the reference case is also seen in Figure 7S8 with the offset of the points. The results here are similar to what has been identified earlier: GISS has unusual offsets for all reactivities and J-O1D; agreement for P-O3 is much better than for L-O3 and L-CH4; four models show the

410 upward curve matching the top-1% parcels; for L-O3 and L-CH4, GFDL-NCAR have a flat scatter of points and miss the upward curve because they reset the q of the data stream. Day-to-day scatter for the top-10% (defined by the 5-day mean) is tested with the UCI model in Figure S69. This one-model synoptic cloud variability has similar scatter to that seen for the more

17

central models (Figure 7S8) including the rapid increase in L-O3 at the top-1% and the much

415    greater scatter in J-NO2.  The year-to-year variability in the top-10% parcels is shown (Figure
       S710) for the UCI model with year 2016 as the reference case (solid
       line) and years 1997 and 2015 as separate models.   The patterns of scatter here are similar to
       but less than the day-to-day (Figure S96), again showing the importance of 5-day averages, and
       identifying the lower limit of scatter at which this diagnostic can discern differences in model
420    chemistry.  Overall, the


       Note that the J-values have their own top-10% parcels, and that is why J-O1D parcels (all in
       uppermost troposphere) havse better agreement than the top-10% across models and much
       smaller scatter than either J-NO2 which are more sensitive to clouds.
425    or the reactivities.  The top J-O1D values are in the upper troposphere (Figure S2) and less
       influenced by clouds; whereas top J-NO2 values occur in cloud fields and even within the
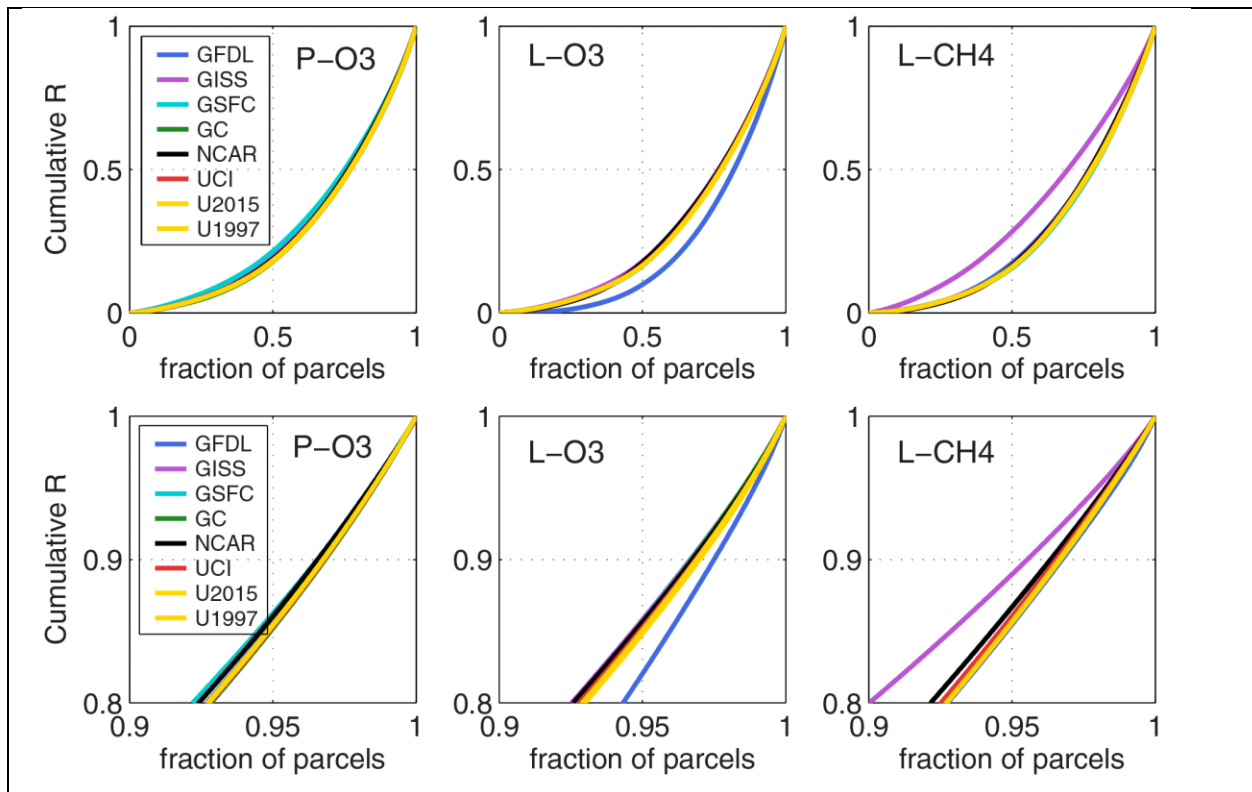       clouds.



Figure 4.  Cumulative reactivity of the 14,880 parcels (equally weighted) scaled to the average of each model and reactivity. The lower panel shows a blowup of the top 20% (Cumulative = 0.8 to 1.0).  Results for the 6 models plus two different years for UCI are shown.
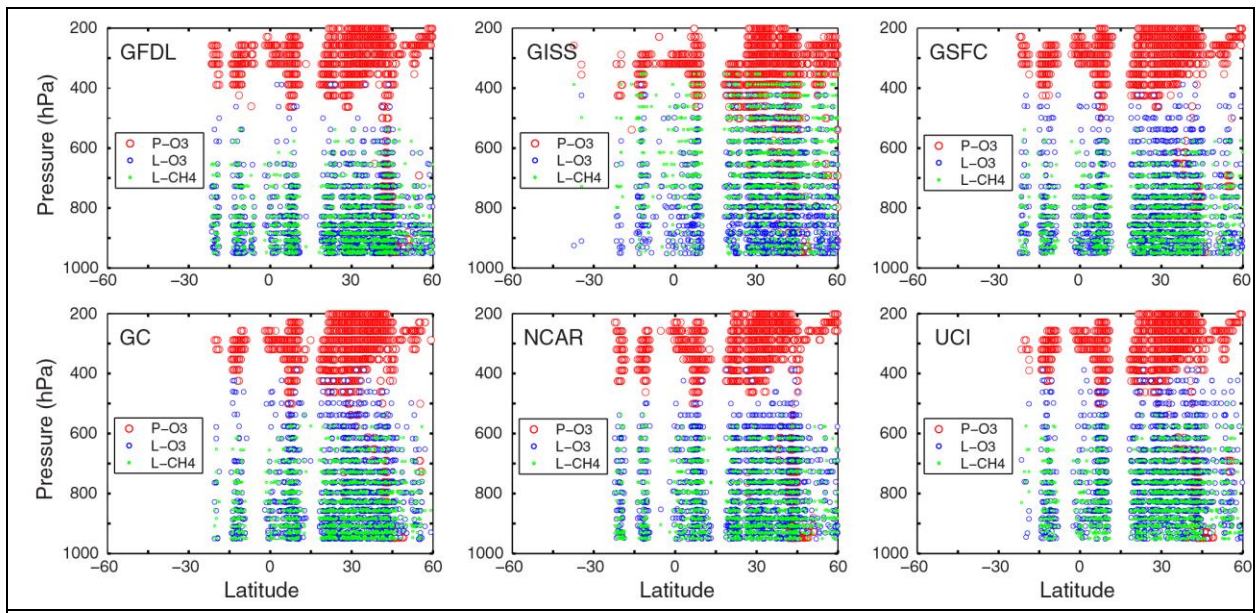
Figure 35. Latitude (degrees) by pressure (hPa) location of the top 10% of reactive parcels for the 6 models: P-O3 (red, large circles); L-O3 (blue, medium); L-CH4 (green, small).
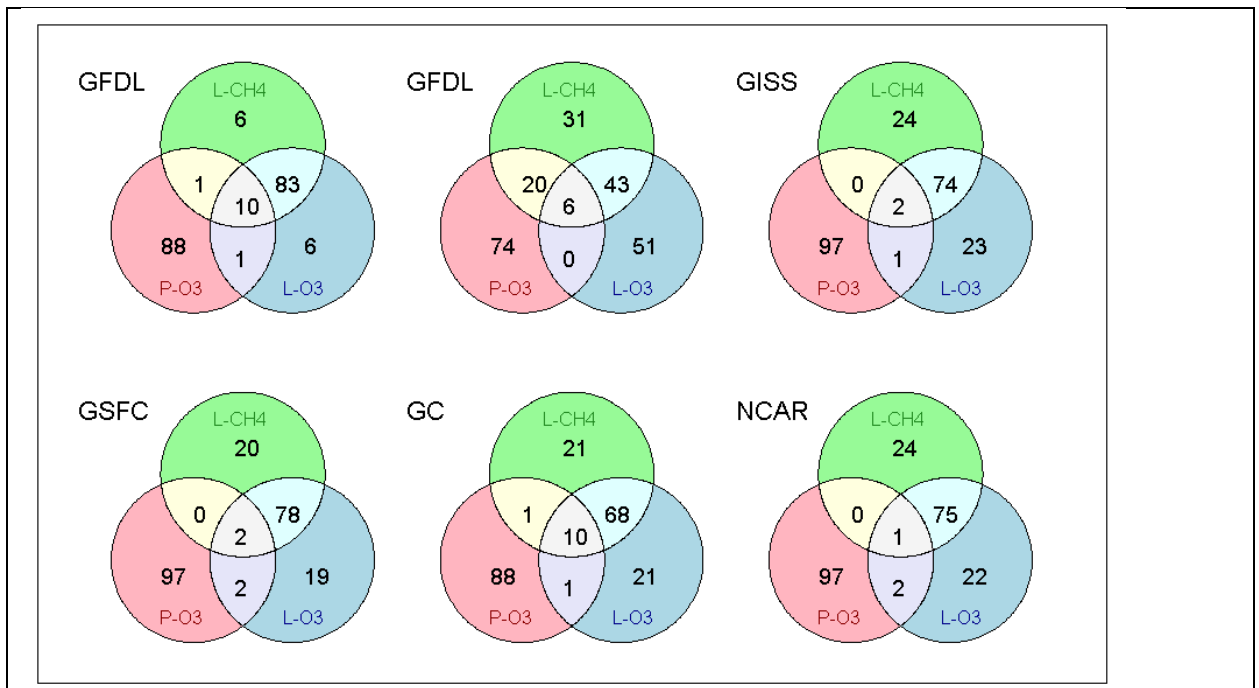
430



Figure 64. Venn diagrams for each model showing the overlap (%) of the top 10% parcels in each reactivity, using 5-day means for each parcel.
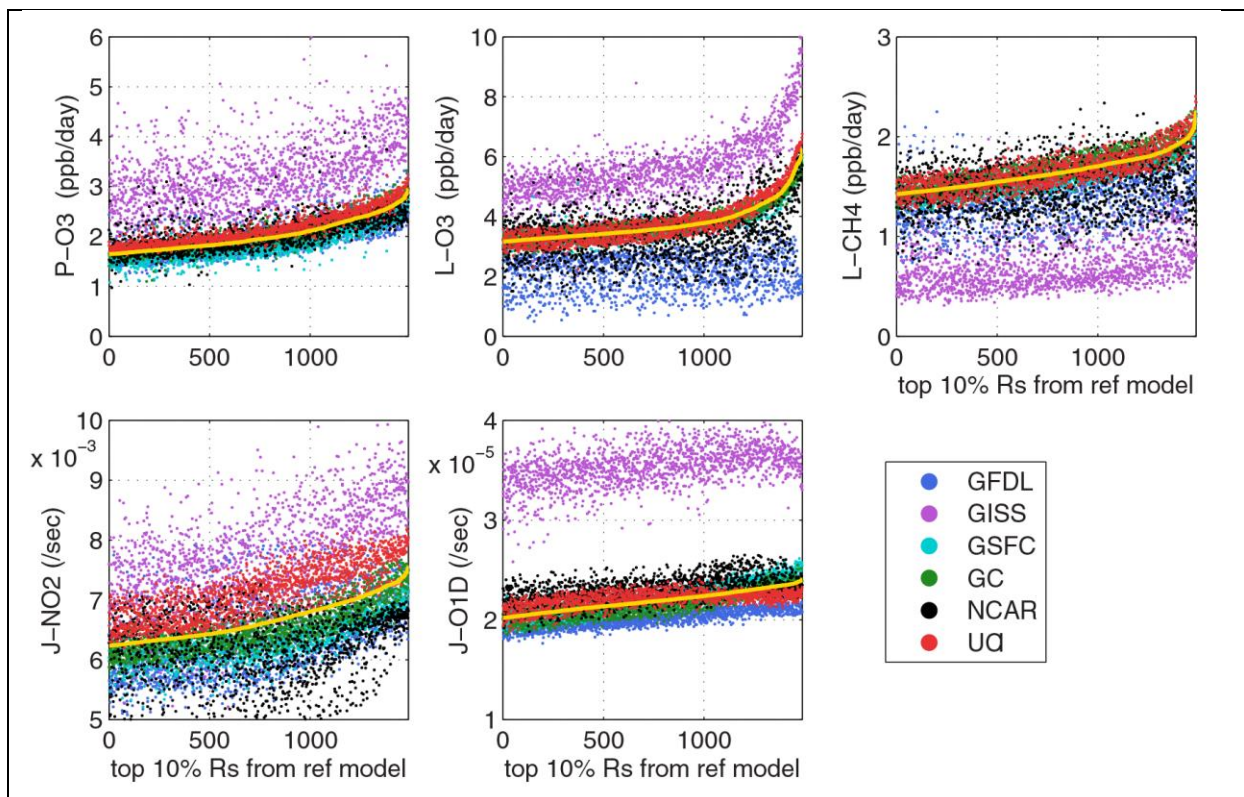
19

Figure 7. Modeled Reactivity and J-values for 5-day mean parcels plotted using the top-10% in the reference case in ascending order along the X axis. The black dashed monotonically increasing line is the reference case parcels.

### 3.5 Assumptions and uncertainties in the experiment design

435

How interannual variability might affect the results is tested with the UCI CTM running the simulated data stream for 5 August days using years 1997 and 2015 meteorology to compare with year 2016 (see previous Tables, Table S89, and Figure 58). The scatter plots in Figure 85 do not look much different from those for the three models used in the reference case (Figure

440    23). For the 5-day parcel means, the rms differences across any pairing of the 3 UCI years is about 8-10% of the average reactivity, which ~~surprisingly~~ is about half of that across the 3 models used in the reference case. Using this criterion (<20%) for distinctness, we effectively have only 4 independent distinctly different models here: GFDL, GISS, NCAR and the GSFC-GC-UCI group. The 4 models all differ from one another at the 30-100% level of the UCI year-

445    to-year variations. In terms of the overall average reactivities (Table 32), however, the different years of the UCI model ~~models~~ are almost identical (<1%), while the differences across the 3 reference models are much larger (±5%) and clearly distinguishable.

20

How the time-of-day of parcels in the data stream might affect reactivity is tested with the UCI

450    model initializing the calculation at midnight (1200H UT) instead of noon (see Figure S118,

Table S89).  In this study, we chose parcels at 180W and, since the global models begin each day

at 0000H UT, the photochemistry starts at local noon.  A measurement data stream, such as from

ATom (2017), will include measurements over a range of longitudes and taken with a wide range

of local solar times.  We need to ensure that the protocol here does not depend on when the 24-

455    hour integration of reactivity is initiated.  The UCI model selected one day (8/16 August 2016)

and shifted the local solar time by 12 hours, thus initiating each parcel at local midnight.  In

addition, the cloud fields needed to be rearranged so that the pairing of clouds and solar zenith

angles were the same in both cases.  The start-at-midnight version has larger reactivities by at

most 1% with no changes in the J-values as expected for the protocol (e.g., keeping the morning

460    clouds in the morning for both calculations).  The rms differences between the two cases are 2-

10 times less than the year-to-year differences.  We conclude that the initiation time produces

discernible differences but not at the level to affect the any of the results here, even with high

levels of lightning-NO in daytime.  The initiation time might affect highly polluted regions

where the NOx reservoirs could be converted at night to less photolabile nitrates.

465

Two additional sensitivity tests included running the 5 days in August with a fixed solar

declination (Figure S129) and with different restart file (Figure S10).  As shown in these figures

Figure S13 and Table S89, these two tests change the overall average in the fourth decimal place

and have rms differences <0.01 ppb/day.  For these choices, the protocol adopted here is

470    adequate.

The GFDL and NCAR CCMs could not maintain the fixed, data-stream T&q values over the 24-

hour integration, which leads to larger rms differences because reactivities depend on both T and

q.  This explains in part why the GFDL and NCAR models in Figure 32 have larger scatter for

475    reactivities than the other non-GISS models, but similar scatter in J values.  This effect may also

contribute to the larger day-to-day rms, for NCAR at least, and is examined more extensively

with the UCI CTM running with the T&q's from both models (Section 3.5).

480    How overwriting of the data stream's T&q (with a CCM climate) impacts these results is tested
       with the UCI CTM re-running a one day (16 August 2016~~8/16~~) data stream using T&q's reported
       ~~out~~ from the GFDL and NCAR models.  The rms reactivity differences for these two models are
       2-3 times larger than those of the reference models (GSFC, GC, UCI, see Table 3~~2~~); while J-
       value~~s~~ differences (much less affected by temperature) are similar.

485

       For the 5 days, each with ~~x~~14,880 parcels, the mean values of either GFDL or NCAR T&q's are
       similar to the data stream but their rms differences are large:  about 3.6ºK and 0.4 in $\log_{10}(q)$, see
       Table S9~~10~~.  Both models have similar scatter patterns for T and for q (Figure S14~~1~~) with a
       number of parcels having $\log_{10}(q)$ more than a factor of 10 different from the stream.  In this
490    sensitivity test, UCI CTM ran with just T from GFDL and NCAR, and then with both T&q (4
       cases in all).  The results are shown in Tables S8~~9~~ and Figure 6~~9~~. For T alone, the reactivity
       differences were at the lower limit of detectable model-model differences but, with both T and q,
       the model showed surprisingly large shifts in L-O3 and L-CH4 along with standard deviations 2-
       10 times larger than the lower limit based on different UCI model years.  In fact, the UCI model
495    using GFDL and NCAR T&q has about the same rms reactivity differences with respect to the
       reference case as do the full models (Compare Tables S8~~9~~ and 3~~2~~, noting that Table 3~~2~~ is a 5-day
       mean result and not 1-~~one~~ day result).  Thus, without a model being able to use the specified
       T&q, we are unable to determine if its photochemical module is similar to another model.
       Moreover, with climate-varying T&q's the modeled reactivities from an observed data stream
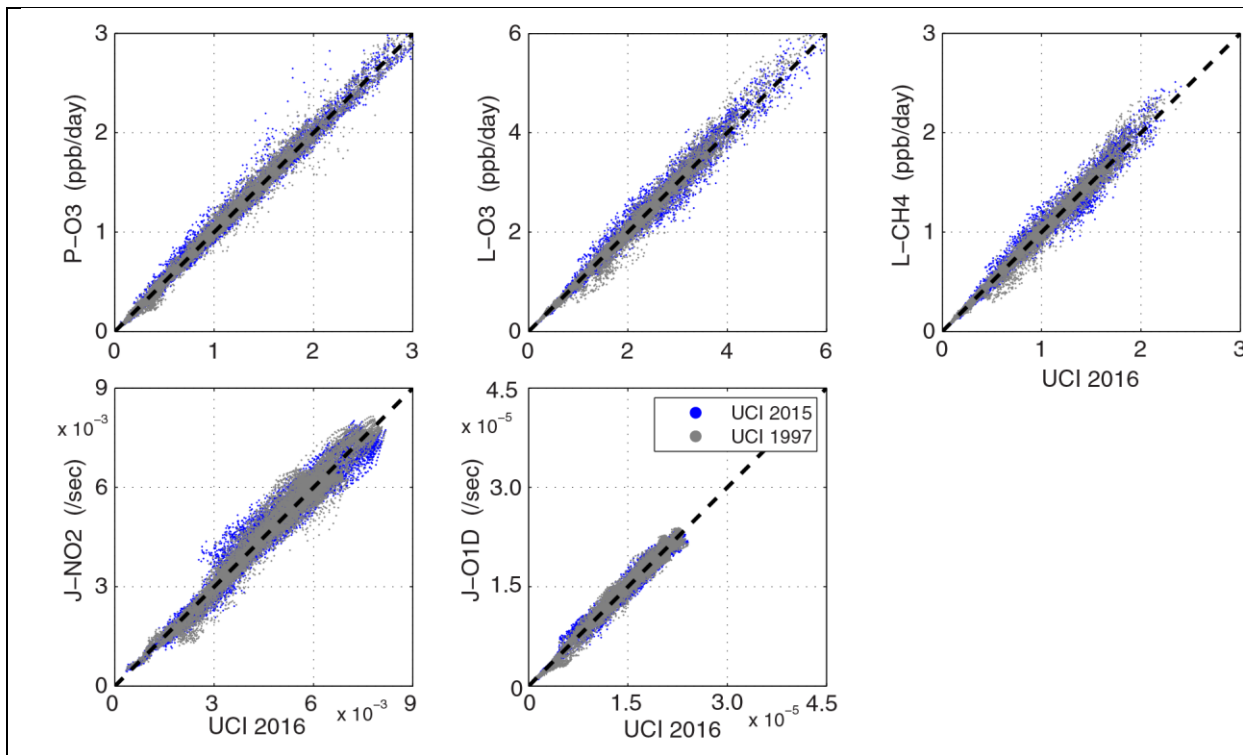500    will also be too noisy for an analysis of the top-10% parcels, i.e., which air matters.

Figure 85. Scatter plot of reactivities and J-values for 5d-mean air parcels for UCI alternate meteorological years (2015, 1997) against the standard year 2016.
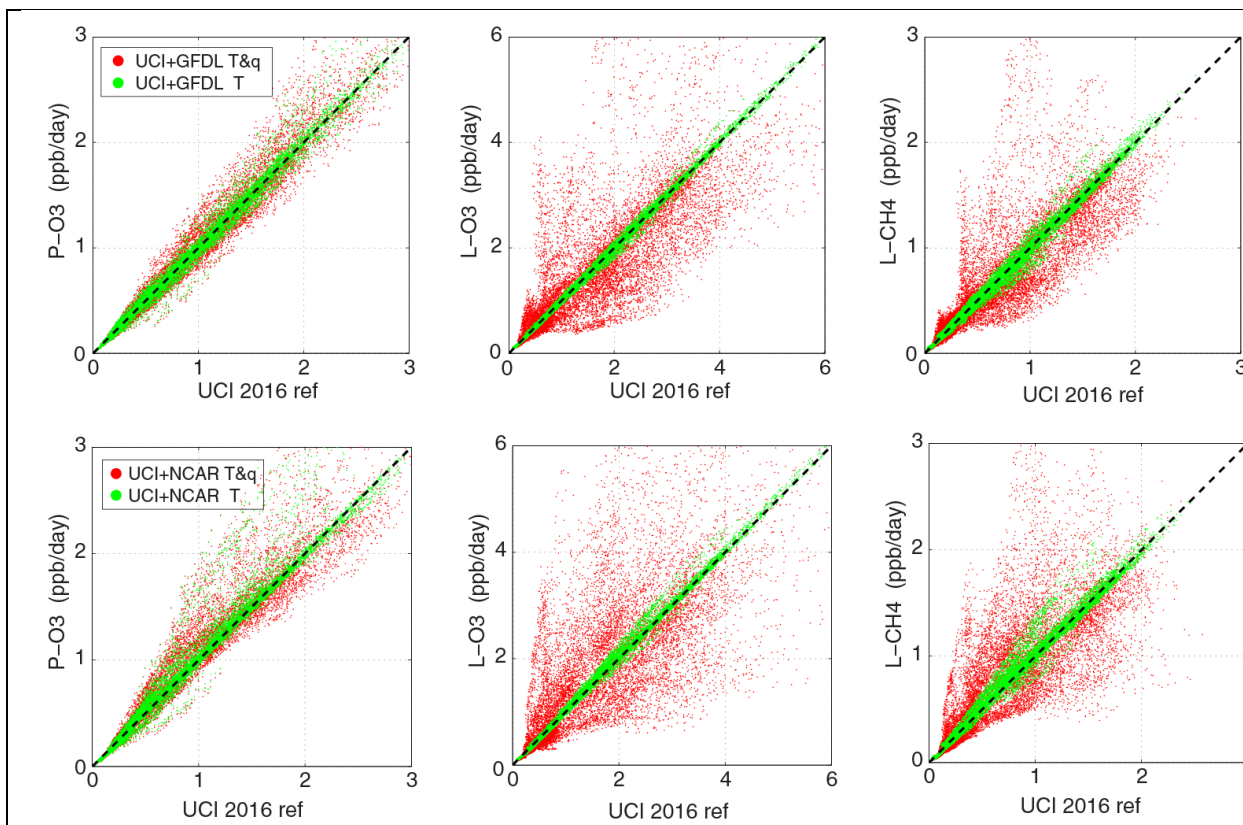


23

Figure ~~6~~9.  Scatter plot of reactivities of the 14,880 air parcels showing the ~~impact~~ effect of the GFDL (top) and NCAR (bottom) models us~~ing~~~~e of~~ T & q ~~their climate models' T (green) and T&q (red)~~ from their climate models, instead of from the specified data stream.  The UCI model for a single day (16 August 2016)~~a single day (8/16 2016)~~ calculated reactivities using the GFDL and NCAR parcel values for T alone (green) and for T&q (red) and compared with the UCI reference model for 16 August 2016. ~~with both prescribed and these models' T&q values.~~

## 4. Summary discussion

We develop a new protocol for merging in situ measurements with 3-D model simulations of atmospheric chemistry as calculated by chemistry-transport models through to Earth system models.  The goal is to take a time stream of species-rich, high-resolution (100s m), spatially sparse observations, such as from an aircraft mission (e.g., ATom, 2017), and have the current 3-D global or regional models use that observed data directly to evaluate chemical reactivity in each parcel.  With this protocol, we avoid model artifacts in the data stream, such as occur in assimilated data, but must account for the density and bias in sampling.  Here, w~~W~~e focus on tropospheric production and loss of the greenhouse gases ozone and methane, but the protocol can be readily applied to other chemical transformations such as the formation and growth of secondary organic aerosols.

In applying the protocol here to a synthetic data stream, we demonstrate a second major use: detailed diagnostics of model performance, specific to the photochemical modules operating within the global chemistry-climate and chemistry-transport models.  Six such models are evaluated here, and their differences and similarities in simulating the chemistry are clearly identified.  The protocol specifies the detailed chemical composition of a constrained set of air parcels including temperature and water vapor, embeds these parcels in an appropriate grid cell of each model, turns off processes that mix adjacent grid cells, and integrates the 3-D model for 24 hours (see P2017).  The photochemical module is thus dependent only on the chemical mechanism and the diurnal cycle of photolysis rates, which are driven in turn by temperature, water vapor, solar zenith angle, clouds, possibly aerosols and overhead ozone, which are calculated as they would be in each model.

Typical 3-D multi-model evaluations cannot separate differences in photochemistry from differences in emissions, transport, scavenging, and even numerical methods, all of which help define the mix of chemical species in each grid cell.  The new protocol established in this paper

535    combines the no-transport A-run from P2017 with the data stream of specified-composition air parcels.  The approach is generic and can be implemented in any model.  Here, using six global chemistry-transport or chemistry-climate models, we can see how it opens a window focusing specifically on the photochemical modules embedded in 3-D models.

540    Overall, the models show surprisingly good agreement on calculating the reactivity (P-O3, L-O3, L-CH4) and photolysis rates (J-NO2, J-O1D) in air parcels.  We can identify unique features in each model:  e.g., UCI's high J-NO2 values; GSFC's lower P-O3 at high reactivity; GISS's inverted results for L-O3 versus L-CH4; GFDL and NCAR's large scatter due to use of model-generated versus parcel-specified water vapor; and large variability in J values for NCAR and

545    GISS.  Models with effectively the same chemistry module will appear distinct if they use a different data stream for water vapor.  It is impossible to tell if overall, among the six models, GISS has the most unique features, and GC the least.  These anomalous features can really only be explained by the model developers who understand the coding, yet these diagnostics point to a focus for the analysis of individual models.  Being a standout in any diagnostic, does not

550    necessarily imply that uniqueness is an error, but it should encourage self-evaluation to determine if that unique feature is intentional and can be shown to be a more accurate simulation.

Cloud variations on synoptic scales are primary sources of noise in this study.  These are difficult to standardize from either model or observation given the wide range of methods for treating

555    cloud scattering and overlap.  Cloud-driven changes in reactivity are clear in comparisons across models and also within the same model.  Use of a single day for comparison is inadequate.  This protocol selects 5 days across the month to sample cloud fields, and this provides a stable average for identifying model-model difference.  The protocol also makes several simplifying assumptions that may affect results:  the solar declination over the month is fixed at the mid-

560    month value; and the 24-hour integration is always started globally at the same universal time, meaning at different local solar times across the longitudes.  These issues were tested with a

single model and found to be unimportant compared with the synoptic variability in clouds and other model-model differences.

565    Using day-to-day and year-to-year variability in a single model, we can define a lower limit to the differences, which is essentially the noise in this protocol, such that models are not distinguishably different.  For the most part, we find that the GSFC, GC and UCI models fall into this indistinguishable-from-one-another class because their differences are within a factor of 2 of the estimated noise level.  This grouping may be explained in part by the common heritage of
570    GSFC and GC's tropospheric chemical model, but UCI's chemical mechanism is completely different and much abbreviated.  All other model pairings show much larger differences.

All models agree that the more highly reactive parcels dominate the chemistry; for example, the hottest 10% of parcels control 25-30% of the total reactivities.  Unfortunately, they do not agree
575    on which parcels comprise the top 10%.  This diagnostic will become more acute as we move from the smoothed synthetic data stream derived from model output (50 x 50 x 1 km averages) to the high variability of in situ ATom observations (2 x 2 x ~0.1 km averages).

Based on our experience comparing models that differ largely by temperature and water vapor,
580    we conclude that water vapor differences in CCM simulations of past and future atmospheres may be a major cause of the changes in $O_3$ and $CH_4$ and may lead to different chemistry-climate feedbacks across the models.

This new protocol for multi-model evaluations helps identify and provide insights into inter-
585    model differences, as well as providing for a direct link with measurements made at a much finer scale than the models.

## References

595

ATom, Measurements and modeling results from the NASA Atmospheric Tomography Mission, https://espoarchive.nasa.gov/archive/browse/atom, doi: 10.5067/Aircraft/ATom/TraceGas_Aerosol_Global_Distribution, 2017.

600 Batchelor, G. K.: The Effect of Homogeneous Turbulence on Material Lines and Surfaces, Proc R Soc Lon Ser-A, 213, 349-&, DOI 10.1098/rspa.1952.0130, 1952.

Brasseur, G. P., D. A. Hauglustaine, S. Walters, P. J. Rasch, J.-F. Müller, C. Granier, and X. X. Tie: MOZART, a global chemical transport model for ozone and related chemical tracers: 1. Model description, J. Geophys. Res., 103(D21), 28265–28289, doi:10.1029/98JD02397, 1998.

605 Collins, W. J., Lamarque, J. F., Schulz, M., Boucher, O., Eyring, V., Hegglin, M. I., Maycock, A., Myhre, G., Prather, M., Shindell, D., and Smith, S. J.: AerChemMIP: quantifying the effects of chemistry and aerosols in CMIP6, Geosci Model Dev, 10, 585-607, 10.5194/gmd-10-585-2017, 2017.

Duncan, B.N., S.E. Strahan, Y. Yoshida, S.D. Steenrod, and N. Livesey: Model study of cross-tropopause transport of biomass burning pollution, Atmos. Chem. Phys., 7, 3713-3736, 2007.

610 Eyring, V., et al.: Assessment of temperature, trace species, and ozone in chemistry-climate model simulations of the recent past, J Geophys Res-Atmos, 111, -, Artn D22308 Doi 10.1029/2006jd007327, 2006.

Gelaro, R. et al.: The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2), J Climate, 30(14), 5419–5454, doi:10.1175/JCLI-D-16-0758.1, 2017

Holmes, C.D., M. J. Prather, A.O. Søvde, G. Myhre: Future methane, hydroxyl, and their uncertainties: key climate
615 and emission parameters for future predictions, Atmos. Chem. Phys., 13, 285–302, doi:10.5194/acp-13-285-2013, 2013.

Horowitz, L. W., et al.: A global simulation of tropospheric ozone and related tracers: Description and evaluation of MOZART, version 2, J. Geophys. Res., 108(D24), 4784, doi:10.1029/2002JD002853, 2003.

Houweling, S., Dentener, F., and Lelieveld, J.: The impact of non- methane hydrocarbon compounds on
620 tropospheric photochemistry, J. Geophys. Res., 103, 10673–10696, 1998.

Lauritzen, P. H., Ullrich, P. A., Jablonowski, C., Bosler, P. A., Calhoun, D., Conley, A. J., Enomoto, T., Dong, L., Dubey, S., Guba, O., Hansen, A. B., Kaas, E., Kent, J., Lamarque, J. F., Prather, M. J., Reinert, D., Shashkin, V. V., Skamarock, W. C., Sorensen, B., Taylor, M. A., and Tolstykh, M. A.: A standard test case suite for two-dimensional linear transport on the sphere: results from a collection of state-of-the-art schemes, Geosci Model
625 Dev, 7, 105-145, 10.5194/gmd-7-105-2014, 2014.

Li, J., Mao, J., Fiore, A. M., Cohen, R. C., Crounse, J. D., Teng, A. P., Wennberg, P. O., Lee, B. H., Lopez-Hilfiker, F. D., Thornton, J. A., Peischl, J., Pollack, I. B., Ryerson, T. B., Veres, P., Roberts, J. M., Neuman, J. A., Nowak, J. B., Wolfe, G. M., Hanisco, T. F., Fried, A., Singh, H. B., Dibb, J., Paulot, F., and Horowitz, L. W.: Decadal change of summertime reactive nitrogen species and surface ozone over the Southeast United States,
630 Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2017-606, 2017.

Madronich, S.: Photodissociation in the atmosphere: 1. Actinic flux and the effect of ground reflections and clouds, J. Geophys. Res., 92, 9740 – 9752, 1987.

Morgenstern, O., Hegglin, M. I., Rozanov, E., O'Connor, F. M., Abraham, N. L., Akiyoshi, H., Archibald, A. T., Bekki, S., Butchart, N., Chipperfield, M. P., Deushi, M., Dhomse, S. S., Garcia, R. R., Hardiman, S. C.,
635 Horowitz, L. W., Jockel, P., Josse, B., Kinnison, D., Lin, M. Y., Mancini, E., Manyin, M. E., Marchand, M., Marecal, V., Michou, M., Oman, L. D., Pitari, G., Plummer, D. A., Revell, L. E., Saint-Martin, D., Schofield, R., Stenke, A., Stone, K., Sudo, K., Tanaka, T. Y., Tilmes, S., Yamashita, Y., Yoshida, K., and Zeng, G.: Review of the global models used within phase 1 of the Chemistry-Climate Model Initiative (CCMI), Geosci Model Dev, 10, 639-671, 10.5194/gmd-10-639-2017, 2017.

640 Myhre, G., Aas, W., Cherian, R., Collins, W., Faluvegi, G., Flanner, M., Forster, P., Hodnebrog, O., Klimont, Z., Lund, M. T., Mulmenstadt, J., Myhre, C. L., Olivie, D., Prather, M., Quaas, J., Samset, B. H., Schnell, J. L., Schulz, M., Shindell, D., Skeie, R. B., Takemura, T., and Tsyro, S.: Multi-model simulations of aerosol and ozone radiative forcing due to anthropogenic emission changes during the period 1990-2015, Atmos Chem Phys, 17, 2709-2720, 10.5194/acp-17-2709-2017, 2017.

645 NAP, Causes and Effects of Changes in Stratospheric Ozone: Update 1983, ISBN 0-309-03443-4, National Academy Press, Washington DC, 1984.

NASA, Report of the 1992 Stratospheric Models and Measurements Workshop, (Prather, M.J. and E.E. Remsberg, eds.), Satellite Beach, FL, February 1992, NASA Ref. Publ. 1292, 144+268+352 pp., 1993.

Nault, B. A., Garland, C., Wooldridge, P. J., Brune, W. H., Campuzano-Jost, P., Crounse, J. D., Day, D. A., Dibb, J., Hall, S. R., Huey, L. G., Jimenez, J. L., Liu, X. X., Mao, J. Q., Mikoviny, T., Peischl, J., Pollack, I. B., Ren, X. R., Ryerson, T. B., Scheuer, E., Ullmann, K., Wennberg, P. O., Wisthaler, A., Zhang, L., and Cohen, R. C.: Observational Constraints on the Oxidation of NOx in the Upper Troposphere, J Phys Chem A, 120, 1468-1478, 10.1021/acs.jpca.5b07824, 2016.

Olson, J. R., Crawford, J. H., Brune, W., Mao, J., Ren, X., Fried, A., Anderson, B., Apel, E., Beaver, M., Blake, D., Chen, G., Crounse, J., Dibb, J., Diskin, G., Hall, S. R., Huey, L. G., Knapp, D., Richter, D., Riemer, D., Clair, J. S., Ullmann, K., Walega, J., Weibring, P., Weinheimer, A., Wennberg, P., and Wisthaler, A.: An analysis of fast photochemistry over high northern latitudes during spring and summer using in-situ observations from ARCTAS and TOPSE, Atmos Chem Phys, 12, 6799-6825, 10.5194/acp-12-6799-2012, 2012.

Olson, J., Prather, M., Berntsen, T., Carmichael, G., Chatfield, R., Connell, P., Derwent, R., Horowitz, L., Jin, S. X., Kanakidou, M., Kasibhatla, P., Kotamarthi, R., Kuhn, M., Law, K., Penner, J., Perliski, L., Sillman, S., Stordal, F., Thompson, A., and Wild, O.: Results from the Intergovernmental Panel on Climatic Change Photochemical Model Intercomparison (PhotoComp), J Geophys Res-Atmos, 102, 5979-5991, 1997.

Orbe, C., Waugh, D. W., Newman, P. A., and Steenrod, S.: The Transit-Time Distribution from the Northern Hemisphere Midlatitude Surface, J Atmos Sci, 73, 3785-3802, 10.1175/Jas-D-15-0289.1, 2016.

PhotoComp: Chapter 6 - Stratospheric Chemistry SPARC Report No. 5 on the Evaluation of Chemistry-Climate Models 194-202, 2010.

Prather, M. J., Zhu, X., Flynn, C. M., Strode, S. A., Rodriguez, J. M., Steenrod, S. D., Liu, J. H., Lamarque, J. F., Fiore, A. M., Horowitz, L. W., Mao, J. Q., Murray, L. T., Shindell, D. T., and Wofsy, S. C.: Global atmospheric chemistry - which air matters, Atmos Chem Phys, 17, 9081-9102, 10.5194/acp-17-9081-2017, 2017.

Prather, M. J., Zhu, X., Strahan, S. E., Steenrod, S. D., and Rodriguez, J. M.: Quantifying errors in trace species transport modeling, P Natl Acad Sci USA, 105, 19617-19621, DOI 10.1073/pnas.0806541106, 2008.

Prather, M. J.: Photolysis rates in correlated overlapping cloud fields: Cloud-J 7.3c, Geosci Model Dev, 8, 2587-2595, 10.5194/gmd-8-2587-2015, 2015.

Prather, M., and Jaffe, A. H.: Global Impact of the Antarctic Ozone Hole - Chemical Propagation, J Geophys Res-Atmos, 95, 3473-3492, 1990.

Prather, M.J., J. Hsu (2010), Coupling of nitrous oxide and methane by global atmospheric chemistry, Science, 330: 952-954.

Rienecker, M. M. et al.: MERRA: NASA's Modern-Era Retrospective Analysis for Research and Applications, J Climate, 24(14), 3624–3648, doi:10.1175/JCLI-D-11-00015.1, 2011.

Shindell, D. T. et al.:Interactive ozone and methane chemistry in GISS-E2 historical and future climate simulations, Atmos Chem Phys, 13(5), 2653–2689, doi:10.5194/acp-13-2653-2013, 2013.

Shindell, D., Kuylenstierna, J. C. I., Vignati, E., van Dingenen, R., Amann, M., Klimont, Z., Anenberg, S. C., Muller, N., Janssens-Maenhout, G., Raes, F., Schwartz, J., Faluvegi, G., Pozzoli, L., Kupiainen, K., Hoglund-Isaksson, L., Emberson, L., Streets, D., Ramanathan, V., Hicks, K., Oanh, N. T. K., Milly, G., Williams, M., Demkine, V., and Fowler, D.: Simultaneously Mitigating Near-Term Climate Change and Improving Human Health and Food Security, Science, 335, 183-189, DOI 10.1126/science.1210026, 2012.

Strahan, S.E., A.R. Douglass, and P.A. Newman: The contributions of chemistry and transport to low Arctic ozone in March 2011 derived from Aura MLS Observations, J. Geophys. Res., 118, doi:10.1002/jgrd.50181, 2013.

Tilmes, S., Lamarque, J.-F., Emmons, L. K., Kinnison, D. E., Marsh, D., Garcia, R. R., Smith, A. K., Neely, R. R., Conley, A., Vitt, F., Val Martin, M., Tanimoto, H., Simpson, I., Blake, D. R., and Blake, N.: Representation of the Community Earth System Model (CESM1) CAM4-chem within the Chemistry-Climate Model Initiative (CCMI), Geosci. Model Dev., 9, 1853-1890, https://doi.org/10.5194/gmd-9-1853-2016, 2016.

Wild, O., Zhu, X., and Prather, M. J.: Fast-J: Accurate simulation of in- and below-cloud photolysis in tropospheric chemical models, J Atmos Chem, 37, 245-282, 2000.