Response to Reviewer #2's comments

This paper represents a huge task, assembling and comparing the results from the multi-model HTAP2 study. It is a brave undertaking. However, I do have some concerns about what was learned in the process. I believe the stated goals of the paper are not well met or met in a cursory fashion. There are a number of inferences stated as fact but not in fact proved. In some cases more analysis seems to be needed. In other cases a clearer explication of what has been learned would be helpful. Recommendations about future work are succinctly summarized, but the paper needs to be stronger in detailing what was learned and in justifying the methodology used.

We thank the careful review by Reviewer #2. Please see below our response (in blue) to the general and specific comments (in black). Additional results and discussions have been added to the text to clarify the methodology and help strengthen the key points.

Major Comments:

I) The stated paper goals are to address: "1) the differences in O3 sensitivities generated from the HTAP2 and HTAP1 experiments to help address how the LRT impacts on NAM changed through time; 2) how the multi-model approach, as well as the refined model experiment design in HTAP2 can help advance our understanding of the LRT impacts, especially the benefits of increasing the global models' resolutions and involving the regional models; 3) the usefulness of satellite observations for better understanding the sources of uncertainties in the modeled total O3 (e.g., from the emission and regional models' boundary condition inputs) as well as for reducing the uncertainties in some of these model inputs via chemical data assimilation." As the paper stands it is not clear if it achieved its goals. The answers to these questions should be clearly articulated in the conclusions and in the body of the paper itself. In particular: 1) Between HTAP1 and HTAP2 models have changed, emissions have changed and the transport has changed. So it is not really clear how the sensitivity changed through time. The authors suggest many of the changes are due to the changes in emissions, but this remains to be proven. The authors could determine if the changes in the sensitivities are consistent with the change in emissions by using the HTAP1 emissions and the current sensitivities (delta O3/delta emissions) to determine if most of the changes from HTAP1 are consistent with emission changes. However, as it stands the first goal of the paper cannot be met without substantially more analysis.

The comparisons of HTAP1 and HTAP2 findings over larger spatial/temporal scales in this paper are limited to the total sensitivities themselves, and disentangling the cause of these changes is beyond the scope of this study. However, rather than simply reporting these differences, we now do have extended discussions to point out that these different sensitivities can be attributed to the following factors:

1) changes in anthropogenic emissions from 2001 to 2010 (HTAP1 to HTAP2)
2) climate variability driven interannual variability of LRT. We now cited the Lin et al. (2014) work as she suggested, in which stronger LRT impact is suggested in 2010.
3) the experimental design, including the different participating models (and even for the models that participated both HTAP1 and HTAP2, different versions and configurations were implemented), SR domain definitions

The objective 2) of this work has been modified to: "how the refined modeling experiment design in HTAP2 can help advance our understanding of the LRT impacts on NAM, particularly the

involvement of regional models and the inclusion of small spatial/temporal scale analysis during high O_3 episodes that are more relevant to air quality management." These also help address your following general comments.

2) It is not clear how this study enhanced our understanding of LRT nor is it very clear how changes in model resolution impacted the solutions. The STEM model resolution is 60x60 km, actually rather comparable to a global model of 1o resolution (about 85 km at 40N). While there is a wide range of different resolutions in the global models it is unclear how this paper really explored the impact of resolution on the results. What aspects of LRT did the paper enhance? This should be clear in the paper.
Please see our response to your general comment II.

3) The usefulness of satellite data is essentially a "motherhood" statement. It is somewhat unclear how this paper further showed this usefulness. This is especially true since the case study using satellite data was presented in a rather cursory manner.
Please see our response to your general comment III regarding the use of satellite data in the case study. The text in the introduction, Section 2 and Section 3.1 explain the purpose, methods, and the findings of using OMI NO_2 data to evaluate the bottom-up emissions. These were also explicitly mentioned in Section 2.3.2, and in the abstract and Section 4 as a highlight in HTAP2.

We also make the readers aware of the uncertainty of the satellite products. For example, for the use of OMI NO_2: "It is important to note that uncertainty in satellite retrievals can prevent us from producing accurate assessment on emissions (e.g., van Noije et al., 2006), and this comparison does not account for the biases in the used OMI data, and would be further validated by using other OMI NO_2 products as well as the bias-corrected (if applicable) in-situ NO_2 measurements." For TES and IASI O_3, "TES O_3 is generally positively biased by <15% relative to high accuracy/precision reference datasets (e.g., Verstraeten et al., 2013). Although IASI is in general less sensitive than TES due to its coarse spectral resolution, the 681–316 hPa partial column-averaged O_3 mixing ratios in the JPL product agree well with TES O_3 for the 2008–2011 period with a -3.9 ppbv offset (Oetjen et al., 2016)."

II) It is not clear what the goal of using the STEM model is here. As pointed out above the resolution is not that much higher than some of the global models that give the boundary conditions. Differences between the STEM results and the boundary condition model could be due to the different chemistry in the two models or due to the differences in transport. Driving the models with different meteorological datasets also risks an inconsistency in the boundary conditions (e.g., chemical plumes transported in the jet in the parent model might be mismatched with the jet in STEM). At any rate the rationale for the use of the STEM model should be clearly articulated. What did we learn by coupling the global models with the STEM model?
As been pointed out in the text and recognized by Reviewers #1 and #3, the use of STEM model here is to test the global-regional model couplings. In Sections 1 and 2, we introduced that "For regional simulations over the North America and Europe, boundary conditions were mostly taken from a single model such as the ECMWF C-IFS or GEOS-Chem.", while in this study we "Extending the HTAP2 regional simulations' basic setup, the STEM top and lateral chemical boundary conditions were downscaled from three global models' (i.e., the Seoul National University (SNU) GEOS-Chem, RAQMS, and the ECMWF C-IFS)". As a key finding of this

work, which is also relevant to your next comment, we did show in case studies that all of the global models performed poorly for for some high O_3 events (except RAQMS with data assimilation). We believe such uncertainty poses difficulties for regional models (regardless of their resolutions and other configurations, parameterization) to accurately estimate total O_3 and the SR relationships using boundary conditions downscaled from these models. This finding provides important information for future regional modeling works on higher resolutions and this point has been sharpened in the revised paper.

Please note that all three global models used to be coupled with STEM are known to have satellite chemical data assimilation capability. Given that satellite assimilation can improve the modeled O_3 performance (as demonstrated in this paper for STEM/RAQMS and in a previous study for STEM/GEOS-Chem), near the end of the paper, we suggested directions for future multi-scale modeling works: "As chemical data assimilation techniques keep developing (Bocquet et al., 2015), several HTAP2 participating global models have already been able to assimilate single- or multi-constitute satellite atmospheric composition data (e.g., Miyazaki et al., 2012; Parrington et al., 2008, 2009; Huang et al., 2015; Inness et al., 2015; Flemming et al., 2017). Comparing the performance of the assimilated fields from different models, and making the global model assimilated chemical fields in the suitable format for being used as boundary conditions would be very beneficial for future regional modeling, as well as for better interpreting the pollutants' distributions especially during the exceptional events…."

We used STEM calculations also because we saved STEM O_3 calculations hourly everywhere within the regional domain, while most of the HTAP2 global models did not do so in all model grids. Using hourly observations is important to generating more accurate MDA8 based analysis and comparing the model fields with satellite observations, which are more policy relevant and are favored components by other reviewers.

While we agree that ideally it'd be better to perform all STEM simulations on a finer resolution grid, that has been determined to be not so practical due to the limitations in time and computational resources, especially that the STEM modeling work shown here is a voluntary/unfunded activity. However, 12 km STEM/RAQMS test simulations were indeed performed and the results have been presented at previous HTAP workshops (e.g., http://www.htap.org/meetings/2015/2015_May_11-15/Powerpoint%20Presentations/Monday/Huang%20HTAP_05112015.pdf). These simulations were not updated to account for the later updates in the HTAP2 emission inventory and are therefore not suitable to be included in this manuscript. However, the findings are overall qualitatively similar to the results based on the 60 km simulations in this paper, for example, STEM/RAQMS and RAQMS show similar spatial patterns and domain-mean values of the sensitivities; STEM/RAQMS free run and RAQMS free run show negative biases (relative to CalNex ozonesonde and aircraft in-situ measurements) in free tropospheric O_3 during high O_3 episodes, which were reduced by satellite data assimilation.

Yes, it is understood that "Driving the models with different meteorological datasets also risks an inconsistency in the boundary conditions (e.g., chemical plumes transported in the jet in the parent model might be mismatched with the jet in STEM)." This does not seem to be a big issue in this study. However, we do think that it is worth carrying out additional experiments in the future to

determine if such inconsistency can be resolved by using the boundary condition models' meteorological fields as WRF's initial and boundary conditions.

III) The case study is rather thin. What are the goals of this section? This section should either be expanded or dropped.
Section 3.3 includes event-based analysis that is more relevant to air quality management than the larger scale results, which is favored by other reviewers, and as a result is an important part of this paper. We have expanded this section and added a summertime case study for comparison as other reviewers suggested. Figures 14/17 evaluate the modeled O_3 vertical distributions during LRT events, showing that all of the global models performed poorly for for these high O_3 events (with the exception of RAQMS with data assimilation). The underestimated "transported background" O_3 levels were connected with the underpredicted surface O_3 exceedances in the western US shown in Figures 15/18. We believe such uncertainty poses difficulties for regional models (regardless of its resolution and other configurations, parameterization) to accurately estimate the total O_3 and the SR relationships using boundary conditions downscaled from these global models.

As other types of observed O_3 vertical profiles, such as ozonesonde data, are not available during one of the events we show and are only available in limited regions (only in California) during another event, we believe that evaluating the boundary condition models using satellite O_3 vertical profiles during the selected high O_3 episodes is new and very informative.

Specific Comments:

1. L42. The sentence beginning is rather awkward. Consider rewording.
Reworded.

2. L48-49, "This indicates. . ...". This has to be proven. As is well known interannual variability of the atmosphere is substantial.
Interannual variability has been included in the discussions, which also addresses the comments by Dr. Lin and Reviewer #3.

3. L175. Starting here the manuscript goes into considerable detail about how the simulatons are set up. This does not work well in the introduction, but belongs in the methodology section.
This paragraph has been substantially modified, with specific goals of the study stated first (also accounting for Reviewer #3's suggestions), and some details of the methods were moved to Section 2.

4. L202, Section 2.1. The manuscript parses the emissions between East Asia, MICS Asian regions and south Asian countries. The domains of each these regions is not clear.
MICS Asia is defined in text as: "MICS-Asia regions (south, southeast, and east Asia, based on country inventory for China and from the Clean Air Policy Support System and the Regional Emission inventory in ASia 2.1, more information also in Li et al., 2017)…" Figure 1 defines the different part of the Asian regions for HTAP2's SR relationship study.

5. Table 1. All abbreviations should be defined. Also the table headings need to be reformatted.
Done.

6. L250-253. This notation is should be improved: the left hand side of the equation has a percentage sign, but not the right. I would suggest something like EASALL(-20%) on the right to distinguish this from the R(O3, EAS, 100%) where presumably all EAS emissions are reduced by 100%.
Done.

7. L290 and following paragraph. A long discussion is presented concerning STEM lightning emissions, biogenic emissions and VOC speciation. How were these emissions parameterized in the other models, the same as STEM or differently? Please specify more thoroughly differences in emissions between STEM and other models.
The non-anthropogenic emissions do differ by models, which impact the background O_3 estimation. See Table 1c, Figure S1 for detailed comparisons between GEOS-Chem and STEM, as well as summary for the boundary condition models. We agree and suggest that for future activities the non-anthropogenic emissions should be formally reported for all models by region and species. We now added in Section 2.1: "Non-anthropogenic emission inputs used in different models' simulations may differ, and their impacts on the modeled total O_3 and the SR relationships will be compared in detail in future studies." And for STEM and its BC models at near L290, we added: "Note that non-anthropogenic emission inputs used in STEM and its boundary condition models differed, as summarized in Table 1c. Figure S1 shows detailed comparisons between STEM and GEOS-Chem's non-anthropogenic (i.e., soil, lightning, biomass burning) NO_x emission inputs, and their impacts on the modeled NAM background O_3 were included in Lapina et al. (2014). Such quantitative comparisons will also be carried out between STEM and its other boundary condition models in future studies."

8. L394. "less sensitive" – less sensitive to what?
To the changes in the "true" state, which can be measured by the averaging kernels. This is introduced by a sentence in the following paragraphs: "A_{TES} is the averaging kernel matrix reflecting the sensitivity of retrieval to changes in the true state (Rodgers, 2000)." Comparison of the TES and IASI sensitivities can be found in Oetjen et al. (2016).

9. L420. "de-stripped" – the meaning is unclear.
Corrected to "de-striped". This is described in Boersma et al. (2011a) which we cited.

10. L459. "suggesting that using". This seems rather speculative. There are many possible explanations.
The discussion has been changed to: As reported in the literature (e.g., Geddes et al., 2016; Travis et al., 2016), the representation of land use/land cover, boundary layer mixing and chemistry can be sources of uncertainty for certain global model (i.e., GEOS-Chem), but how serious these issues were in the other models need to be investigated further. Some other possible reasons include the variation of these models' non-anthropogenic emission inputs and chemical mechanisms (Table 1c). Future work should emphasize on evaluating and comparing all models on process level to better understand their performance. Except in the northeastern US, the eight-model ensembles show better agreement with the CASTNET O_3 observations than the three boundary condition-model ensemble. Overall the three-model ensemble only outperforms one model but the eight-model ensemble outperforms seven. This reflects that averaging the results from a larger number

of models in this case more effectively cancelled out the positive or negative biases from the individual models."

11. L471-472. "overall there does appear to be a positive bias". This seems to be a rather strong statement considering the previous sentence. It would be better to say satellite is consistent with a positive bias.
This sentence has been reworded to: "While grid-scale differences in NO_2 columns may not be directly indicative of emissions biases (Qu et al., 2016), these discrepancies are possibly due to a positive bias in the bottom-up emissions, mainly from the anthropogenic sources, which have also been pointed out by Anderson et al. (2014) and Travis et al. (2016)."

12. L478-479. Can you provide a reference why co-emitted species are likely to be biased in the same way as NOx. It is not at all clear to me that emission factors would be all biased in one direction.
Janssens-Maenhout et al. (2015) and Li et al. (2017) summarized that generally the uncertainty ranges are relatively small for species whose emissions are dominated by large-scale combustion sources but larger for those from small-scale and scattered sources. Based on such information, this part of discussion has been modified to reflect the sector and species dependent uncertainty ranges. Additional text was added to Section 2.1 as well.

13. L509-510. "mainly due to". Maybe. It would be better to say consistent with.
"mainly" was changed to "in part".

14. L556-557. Did you show this? Probably better to say "consistent with".
The literature we cited showed this. This sentence has been changed to "the substantial improvement in the European air quality over the past decades that is shown in Crippa et al. (2016) and Pouliot et al. (2015), which contrasts with the growing anthropogenic emissions from the East Asia and other developing countries during 2001-2010". Discussions were also extended to other reasons causing the differences between HTAP1 and HTAP2.

15. L567-568. This is an interesting result: that R in HTAP2 is larger than in HTAP1. However, the reasons for this have not been clearly shown. Certainly the difference is consistent with emission trends but the authors need to establish that this is the case (see general comments above)
Please see our response to your first general comment.

16. Figure 9. I think this is a scatter plot of R(MDA8,EAS,20%) and R(O3, EAS, 20%). Please address the notation.
Figure 9 in the original submission to ACPD in Oct 2016 was moved to the supplement per Reviewer #1's suggestion.

17. The point of section 3.3 is not clear. Some of the figures panels in this section seem to be referred to in a very cursory manner or not at all (e.g., Figure 11). This section needs to be much better developed or not presented.
Same as the response to your general comment (III): Section 3.3 includes event-based analysis that is more relevant to air quality management than the larger scale results, which is favored by other reviewers, and as a result is an important part of this paper. We have expanded this section and

added a summertime case study for comparison as other reviewers suggested. Figures 14/17 evaluate the modeled O_3 vertical distributions during a LRT event, showing that all of the global models performed poorly for for some high ozone events (with the exception of RAQMS with data assimilation). The underestimated "transported background" O_3 levels were connected with the underpredicted surface O_3 exceedances in the western US shown in Figures 15/18. We believe such uncertainty poses difficulties for regional models (regardless of its resolution and other configurations, parameterization) to accurately estimate the total O_3 and the SR relationships using boundary conditions downscaled from these global models. As other types of observed O_3 vertical profiles, such as ozonesonde data, are not available during one of the events we show and are only available in limited regions (only in California) during another event, we believe that evaluating the boundary condition models using satellite O_3 vertical profiles during selected high O_3 episodes is new and very informative.

18. Figure 7 caption. I assume (a), (b), and (c) refer to the first three rows. Better to say row 1, row2 and row 3 or label all panels with letters.
We have labelled all panels of this figure with letters.