Response to Reviewer #1 (Dr. Tonnesen)'s comments

We thank the careful review by Dr. Tonnesen. Please see below our response (in blue) to her general and specific comments (in black). As a majority of her comments were also received during the ACPD reviewing phase, some changes have been made to the original manuscript to address a number of these comments. The revised ACPD manuscript with tracked changes (submitted in late Nov 2016 to ACP together with a clean version) shows these changes more clearly.

General comments

Most of the paper is focused on comparison of monthly mean model results with very limited evaluation of model performance and no analysis of the causes of the differences among the global model simulations. This analysis is not substantially different from previous HTAP studies and is not informative. I suggest moving most of the text and plots that discuss monthly mean model results to the supplement, and instead, the authors should evaluate and compare model performance on several short term episodes that are more relevant to ozone transport and air quality planning.

The EPA White Paper that you suggested in the later comments summarizes some analyses falling into two categories: 1) the monthly, seasonal, or annual mean analyses that provide a broad characterization perspective. Many published and ongoing analyses with focus on conditions in the past decades are done on such large scales, including HTAP1 and some of the HTAP2 analysis; and 2) those focusing on specific polluted events, which are more important to US air quality management. It is mentioned in the White Paper that as long as the averaging time of the results is clarified, both kinds of analysis would be considered.

A uniqueness of this paper is that it includes analyses on both large and small scales, as now highlighted in the abstract: "In addition to the analyses on large spatial/temporal scales relative to the HTAP1, we also show results on subcontinental- and event-scale that are more relevant to the US air quality management." To meet the objective of this study of connecting results from the past studies, including the HTAP1 and other HTAP2 works, we performed analyses using the multi-model mean approach over large spatial and temporal scales. Moreover, model evaluation over some non-NAM regions would be only possible on a monthly basis (e.g., over East Asia) using the available sparse/infrequent in-situ measurements there. More detailed model evaluation has been added to the paper.

To be more relevant to the US air quality management, we also conducted event-based analysis over the US in May and June 2010, and reported model performance and modeled SR relationships on polluted sites/days. A June event was newly added per your following suggestion. The weight of the O_3 exceedance based analyses in the revised paper significantly increased. See Figures 10-18 and related text.

The most interesting aspect of the paper is the section that addresses the May 9, 2010 O3 episode. I suggest including a more detailed discussion of this event, including an assessment of the relative contributions of stratospheric O3 and international transport of O3 for this event. Given that all of the global models performed poorly for this event (with the exception of RAQMS with data assimilation), a key finding could be that currently available global models do not perform well

for some high ozone events. I also suggest performing additional analysis for at least one other high O3 event during summer 2010 to contrast with the May 9 event. By performing a more detailed evaluation and comparison of the different global models (and the couple STEM/Global model simulations) for specific episodes, the authors can more directly evaluate model performance and the suitability of the individual global models for use as boundary condition data in higher resolution regional models.

Significant changes to the paper have been made to address this:

- We added a summer event (~10 June, 2010) leading to similar conclusions to the existing 9 May case study. See Figures 16-18 and related text.
- Model performance and modeled SR relationships on polluted sites/days are now reported (Figures 10, 11, 12 for May-June 2010; Figures 14/15a-d and 17/18a-d panels for two exceptional events) and the conditions for spring and summer times are compared.
- The impacts of stratospheric O_3 intrusion reported by Lin et al. (2012a, b) for these two events were added to Section 3.3 (i.e., ~1/3 and ~50% of the total at where exceedences occurred based on their model estimates).

We extended the event-based analysis and discussions to highlight the findings from these case studies, for example, as you said, that all of the global models performed poorly for some high O_3 events (with the exception of RAQMS with data assimilation). We believe such uncertainty in the chemical boundary conditions poses difficulties for regional models (regardless of their resolutions and other configurations, parameterizations) to accurately simulate the total O_3 and estimate the SR relationships using boundary conditions downscaled from these global models.

I suggest deleting the text that asserts that the use of an ensemble of global models is a preferred approach. The citation (U.S.EPA 2016) is summary of comments at a public meeting and should not be used as citation because the comments were not peer reviewed and do not reflect the consensus of the meeting participants. A better citation would be the EPA whitepaper on background ozone which was reviewed within EPA and is available at https://www.epa.gov/ozone-pollution/background-ozone-workshopand-information. The whitepaper does not recommend the use of multi-model means to reduce uncertainty. The Li et al. 2016 citation is an analysis of visibility trends and does not evaluate multi-model results. Moreover, there is no valid theoretical basis to assume that the average of poorly performing models will be more accurate than the best performing individual model for key atmospheric processes. While it is possible that, by chance, the average of several poorly performing models will better match observations, the average may still inaccurately represent the individual processes that contribute to O3 and the sensitivity of O3 to emissions reductions. While it might be true that positive and negative bias errors cancel when averaging multiple model results for monthly or seasonal means, this does not necessarily indicate that the multi-model average represents O3 more accurately for episodic events that are of interest to the air quality planning community. A better approach would be to evaluate and compare models at the process level and specifically for high O3 episodes, and then select the best performing individual model.

The EPA White Paper is now cited as "US EPA, 2016a" in the introduction. The Li et al. (2016) citation was removed. Thanks for the comments on the use of multi-model approach, as well as the suggestions on evaluating/comparing the models on process level. A sentence has been added to introduce the multi-model approach: "'Ensemble' model analyses have been suggested by some US stakeholders as one of the methods for helping with the characterization of the background O_3 components (US EPA, 2016b)." The multi-model approach in this paper was mainly used to

connect the findings in HTAP1. We now show individual model's performance in Table 1, Figure 11, 15a-d, 18a-d, and the event-based analysis has been extended in which individual model's performance was shown. The language in the discussions related to the multi-model mean results has been modified. For example, over the US, "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.", but for the East Asia, "Unlike at the CASTNET sites, the three-model ensemble agrees better with the observations than the eight-model ensemble". We listed in this study possible sources of uncertainty for some model and pointed out "Future work should emphasize on evaluating and comparing all models on process level to better understand their performance", which would be good materials for follow-on papers.

Detailed comments

Line 63: Opening sentence is awkward. There is no clear link of the uneven distribution to the health/ecosystem impacts of O3. Also, the uneven distribution of O3 is mostly caused by strong concentration gradients in precursor emissions, but this sentence only identifies the O3 lifetime as a cause of the distribution. Suggest rewriting with a focus on the high mixing ratios, not just the distribution.
The opening sentence was rewritten.

Line 73: "to control the emissions of its precursors from these various sources". Not clear what "these various sources" refers to here. The previous sentence identified the stratosphere and local to distant emissions sources, so presumably this sentence is suggesting that there will be benefits of control of both local and international emissions sources, but this sentence then goes on to list precursors categories (VOC, NOx, CO) without reference to local vs international or biogenic vs. anthropogenic. I can infer what the authors mean, but the introductory paragraphs are awkwardly written and potentially confusing to a reader who is not an expert.
This sentence was rewritten.

Line 75: Also include methane in this list
The original "VOCs" has been split to methane and non-methane VOCs in this sentence.

Line 80: background and baseline are not the same. See Cooper et al. for their definition of baseline, and EPA white paper (link below) for definition of U.S. background ozone. Briefly, baseline O3 (as defined by Cooper et al.) can include contributions from upwind U.S. anthropogenic precursor emissions while U.S. background ozone excludes all U.S. anthropogenic emissions. Line 82: Given how the authors defined baseline/background O3, this statement is problematic: "below which the air quality standard is not recommended to be set". Baseline O3 can be elevated in some areas because of transport of anthropogenic precursors and O3 from upwind U.S. states. It is appropriate to set the NAAQS below the baseline O3 level in these areas because the elevated baseline O3 is being addressed by emissions reductions in upwind states. I recommend breaking this very long sentence into several sentences that describe each of the points identified more clearly and more accurately.
This part has been modified and now reads as: "Issues regarding making accurate estimates of the total O_3 as well as the background O_3 level (defined as the concentration that is not affected by recent locally-emitted or produced anthropogenic pollution) (e.g., McDonald-Buller et al., 2011;

Zhang et al., 2011; Fiore et al., 2014; Huang et al., 2015), have been recently discussed as part of the implementation of the new US O_3 standard (US EPA, 2016a, b)."

Line 90: "It has been revealed" is awkward – "revealed" has other connotations. Suggest "It has been found".
Done.

Line 95: "A better understanding of the processes that determine the O3 distributions" Note that the authors have not yet clearly and comprehensively described the processes. They should describe the roles of stratospheric (both routine contributions and discrete intrusion events), biogenic precursors, wildfires, and anthropogenic precursors. We are especially concerned with conditions in which the mixing ratio exceeds the NAAQS, so it is not only the distribution but also the mixing ratio that is important.
Changed "O_3 distributions" to "O_3 pollution levels…". While those multiple sources contribute to the total O_3 and its exceedances, the component this study mainly focuses on is the LRT of non-NAM anthropogenic pollution, particularly those from the East Asia.

Line 96: delete "for recent years". This will be useful for all past years and for future predictions.
Done.

Lines 110-112: "Large intermodel diversity was found in the simulated total O3 and the intercontinentally transported pollution for the chosen SR pairs in the northern midlatitudes, indicating the challenges with simulations by any individual model to accurately represent the key atmospheric processes." The conclusion that no individual model performs well is not supported by a finding of inter-model diversity. For example, it is possible that one model performs well while other models do not. The authors need to cite results of the individual model performance evaluations to support the statement that no model perform well.
We changed "any individual model" to "model simulations". Now the global models, particularly the three boundary condition models, are evaluated individually in places. Model evaluation at the receptor side (western US) is performed against both the surface in-situ observations and satellite vertical profiles; Evaluation at a focused source region (East Asia) has been added. The model comparison with OMI column data provides the uncertainty introduced by the bottom-up emission inventory. These all help better understand the different models' performance.

Lines 113-116: The citation (U.S.EPA 2016) is summary of comments at a public meeting and should not be used as citation because the comments were not peer reviewed and do not reflect the consensus of the meeting participants. A better citation would be the EPA whitepaper on background ozone which did receive review within EPA and is available at https://www.epa.gov/ozone-pollution/background-ozone-workshop-andinformation. The whitepaper does not recommend the use of multi-model means to reduce uncertainty. The Li et al. 2016 citation is an analysis of visibility trends and does not evaluate multi-model results. Moreover, there is no valid theoretical basis to assume that the average of poorly performing models will be more accurate than the best performing individual model for key atmospheric processes. While it is possible that, by chance, the average of several poorly performing models will better match observations, the average may still inaccurately represent the individual processes that contribute to O3 and the sensitivity of O3 to emissions reductions. While it might be true that positive and

negative bias errors cancel when averaging multiple model results for seasonal or annual means, this does not necessarily indicate that the multi-model average represents O3 more accurately for episodic events that are of interesting to the air quality planning community. A better approach would be to evaluate and compare models at the process level and for high O3 episodes, and then select the best performing individual model.

Same as our response to the last general comment: The EPA White Paper is now cited as "US EPA, 2016a" in the introduction. The Li et al. (2016) citation was removed. Thanks for the comments on the use of multi-model approach, as well as the suggestions on evaluating/comparing the models on process level. A sentence has been added to introduce the multi-model approach: "'Ensemble' model analyses have been suggested by some US stakeholders as one of the methods for helping with the characterization of the background O_3 components (US EPA, 2016b)." The multi-model approach in this paper was mainly used to connect the findings in HTAP1. We now show individual model's performance in Table 1, Figure 11, 15a-d, 18a-d, and the event-based analysis has been extended in which individual model's performance was shown. The language in the discussions related to the multi-model mean results has been modified. For example, over the US, "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.", but for the East Asia, "Unlike at the CASTNET sites, the three-model ensemble agrees better with the observations than the eight-model ensemble". We listed in this study possible sources of uncertainty for some model and pointed out "Future work should emphasize on evaluating and comparing all models on process level to better understand their performance", which would be good materials for follow-on papers.

Lines 123-125: Note that in certain VOC/NOx chemical regimes the model response to NOx emissions can be strongly non-linear for smaller NOx changes, so the statement that 20% emissions were selected to be "small enough in the assumed near-linear atmospheric chemistry regime" is not consistent with how models respond to NOx emissions and does not provide an explanation for using 20% emissions reductions. A 100% reduction in emissions from a source sector or region is a better approach to evaluate source attribution. Lines 126-130 identify problems with the use of a 20% reduction and this should also be noted in the conclusions. For future work, I recommend 100% reductions when evaluating source contributions.

We chose a 20% reduction to be consistent with HTAP1 and HTAP2's experiment design. We cited papers here and also in places in Section 3 comparing the sensitivities in response to different sizes of perturbations and the suitability of each choice for address different questions. We also included a couple of sentences in the conclusion related to the scalability: "…The underestimation in other seasons of the HTAP2 study period may be higher and will need to be quantified in future work. Motivated by Lapina et al. (2014), additional calculations will be conducted in future to explore the scalability of different O_3 metrics in these cases. For future source attribution analysis, in general it is recommended to directly choose the suitable size of the emission perturbation based on the specific questions to address, and to avoid linearly scaling O_3 sensitivities that are based on other amounts of the perturbations."

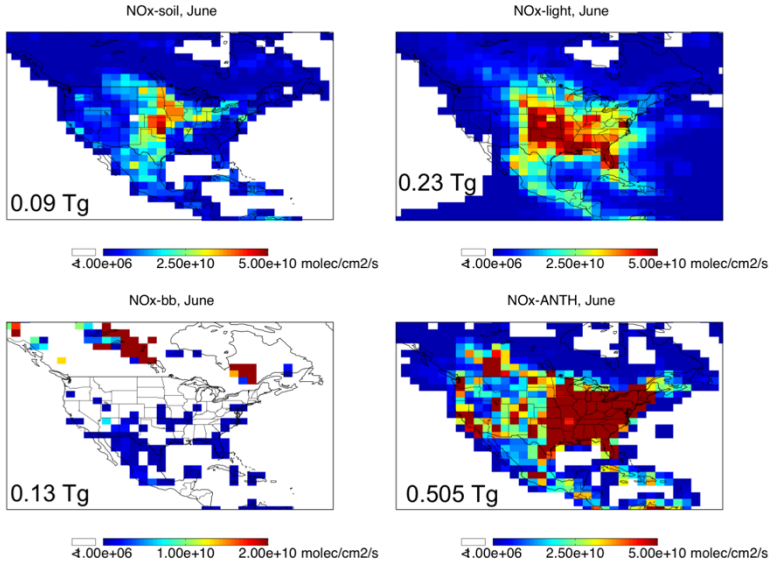
Line 143: "the necessity of evaluating the extra-regional source impacts on event scale [have] has been emphasized" This is a key point – check to see if this addressed in discussion and conclusions.

We have added a summer event case study for comparison. Model performance and modeled SR relationships on polluted sites/days are now reported.

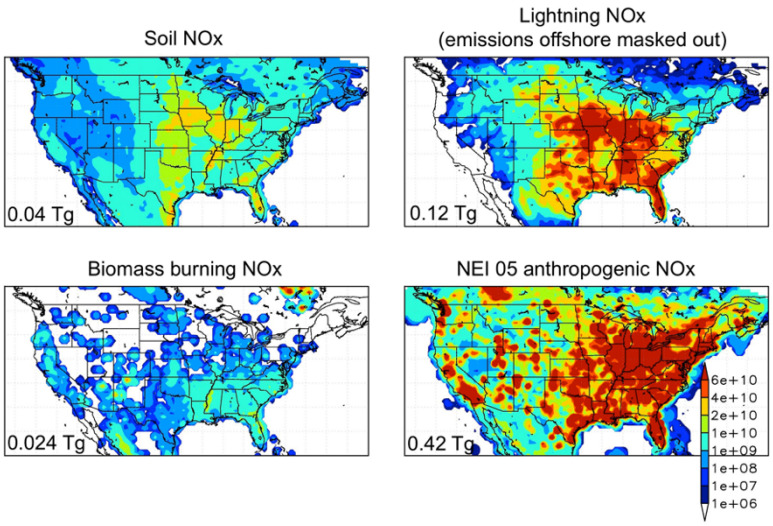
Lines 214-216: Biogenic emissions of VOC are larger than anthropogenic VOC globally, and biogenic and geogenic emissions of NOx, SO2, CO and CH4 are also large and can have a substantial impact on model results. It would have been best to harmonize the natural emissions in addition to anthropogenic emissions, and this approach should be used in future work. For this manuscript, the natural emissions used for each model should be summarized and compared, and, if the natural emissions are significantly different between models, the possible effects on model results should be discussed.

The non-anthropogenic emissions do differ by models, which impact the background O_3 estimation, but these have only been compared in detail between GEOS-Chem and STEM. What's shown in the following plots (also included in the paper discussion and SI) are June 2010 comparisons for soil, lightning, biomass burning and NEI05 anthropogenic NO_x emissions (in molec./cm^2/s) used for the Lapina study, and the numbers at lower-left corners indicate the domain integrated amounts (note that GEOS-Chem emissions were plotted/integrated over a slightly larger domain). The same set of non-anthropogenic emissions was used for our HTAP2 simulations.

GEOS-Chem:



STEM:



Comparing this study's GEOS-Chem emissions with previous studies on summer 2005 (Choi et al., 2009: Soil: 0.05 Tg, lightning: 0.19 Tg; biomass burning: 0.005 Tg; anthropogenic: 0.46 Tg), it seems that non-anthropogenic emissions contributed more to the total CONUS NO_x emissions in June 2010.

GEOS-Chem, C-IFS and WRF/STEM BVOC emissions were all calculated using MEGAN, but the meteorological inputs for their calculations are different (listed in Table 1c), which could lead to notable differences. Wolfe et al. (2015) showed that GEOS-Chem isoprene emissions are ~40% higher than aircraft flux observations in some US regions, and a detail quantification of WRF/MEGAN's biases is included in Huang et al. (2017).

We agree and suggest that for future activities the non-anthropogenic emissions should be formally reported for all models by region, sector, and species. In this section, we now added: "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 boundary condition models, 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."

Line 254: Equation 2 is confusing because the labels for the scenarios are confusing. It is not clear what RERER(O3,NAM) represents. Does this represent a percent contribution from local versus non-local sources?
Equation 2 has been rewritten. For further explanation, a sentence was added following the equation: "The denominator and numerator terms of RERER represent the impacts of global and non-NAM anthropogenic emissions on NAM O_3, respectively."

Lines 227-240: The description of the model scenarios and the naming convention is complicated and difficult to understand. In line 231, why is "all" enclosed parentheses? Why is a 20% sensitivity simulation described as "*source region*ALL". It is not clear what "ALL" means, and generally, the approach used to label the scenarios is not intuitive.
A sentence has been added: "where "ALL" refers to "all species and sectors", consistent with HTAP1 and HTAP2's naming convention."

Line 266-270: Why would lower than normal temperatures in the western U.S. favor decomposition of transported PAN? Lower temperatures would make PAN more stable.
The sentence now reads as: "The mean near-surface air temperatures in the western US in this spring were lower than the climatology, with larger anomalies in the mountain states, which may have led to weaker local O_3 production and decomposition of the transported peroxyacyl nitrates (PAN)."

Lines 287-291: The discussion/conclusions should address the uncertainty introduced by using monthly mean emissions.

Following this sentence, we added a sentence: "This change can introduce uncertainty for some US regions where weekday-weekend variability of some O_3 precursors' emissions was notable during the studied period (e.g., weekend NO_x emissions in southern California during spring/summer 2010 were 0.6-0.7 of the weekday emissions as reported by Kim et al. (2016) and Brioude et al. (2013)), but this was done to ensure consistency with the HTAP2 global model simulations, that also didn't use daily variable emissions for any regions in the world." In Section 3.1.1, we added another sentence: "Also, the use of monthly-mean anthropogenic emissions as well as the overall rough treatment of emission height and temporal profiles can be sources of uncertainty." In conclusion, we now have: "..efforts should also be placed to have the models timely update the height and temporal profiles of the emissions from various sectors". This includes both diurnal and weekly cycles.

Lines 293-294: I doubt that the speciation of VOC emissions in 2005 is substantially different compared to 2010, but if the authors' statement that it is "highly unrealistic" to approximate 2010 using 2005 VOC speciation, this seems to be a significant problem for interpreting the model results.

We agree that the uncertainty of VOC speciation may be high for its base year of 2005 as well. This sentence has been changed to: "The VOC speciation based on the year of 2005 can be unrealistic for 2005 as well as 2010…".

Line 404: Table 2. The model performance evaluation results in Table 2 are not adequate to evaluate the models. In addition to showing the mean bias for multiple models, the model evaluation should also show the bias and error for each model, and the bias and error for the highest observed O3 days because these are the days that are most relevant for air quality planning.

We have done extra work to evaluate the boundary condition model in greater detail. Model performance and modeled SR relationships on polluted sites/days are now reported for the entire study period and during two case studies. In addition to the evaluation over the US, we added the evaluation over the East Asia with the EANET surface observations.

Lines 430-432: "Except in the northeastern US, the eight-model ensembles show better agreement with the CASTNET O3 observations than the three boundary condition model ensemble, suggesting that using a larger number of models in the ensemble calculations may result in better overall model performance." Given that the goal of this study is to evaluate the contributions of international emissions to O3 transport in different regions of the world, it is critically important to understand the individual performance of each global model. If there are substantial difference among models in the contributions of stratospheric O3, chemical production of O3 from precursors, or transport and dispersion of O3, the effect of averaging multiple models may be to introduce additional error into the analysis. A better approach is to compare each global model at the process level, and select the best performing models. If it is uncertain which model performs best, source response relationship should be evaluated using simulation with each BC from each of the global models to estimate the uncertainty in the SR relationships.

This part has been modified 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." We now evaluate each of the global models individually, with strong focus on the three boundary condition models.

Lines 461-463: Recommend showing the individual model performance results using each global model BC instead of averaging the results for all three simulations.
Done.

Table 4: These results are interesting, but to be policy relevant, we need estimates of the contributions on days that exceed the O3 NAAQS. For example, international transport contributions might be highest on days with good dispersion conditions that do not exceed the NAAQS, and lower for days with stagnant dispersion conditions that are more likely to exceed the NAAQS in urban areas. Alternatively, if might be possible that NAAQS exceedances are more likely to occur in rural areas as a result of international transport because of strong mixing from the troposphere to the surface. It is very difficult to interpret the significance of results that are presented as the mean for all days.
As mentioned previously, the monthly-based analyses in this paper were mainly used to connect the findings in HTAP1, and the analysis focusing on polluted sites/days has been extended.

Lines 467-469: This is a key uncertainty that the study does not address. If the modeling systems is biased low for international transport and biased high for local O3 production, the results of the SR analysis may not be reliable.
We point out the uncertainty from the free running HTAP2 simulations in these sentences, but do also suggest the methods to reduce this uncertainty. In the following sentences: "Switching the STEM chemical boundary conditions to the assimilated RAQMS base simulation led to increases in the simulated surface O_3 concentrations by >9 ppbv in the western US (Figure S2, right), associated with higher positive biases (due to several factors discussed in the next paragraph). Regional-scale assimilation could further reduce uncertainties introduced from regional meteorological and emission inputs to obtain better modeled total O_3 and the partitioning of trans-boundary versus US contributions (e.g., Huang et al., 2015)." Additionally, in the last paragraph of this paper, we proposed the possible approaches to improve source attribution estimates by incorporating observations.

Also, the quality of the model boundary conditions only indicates how well the total "transported background" component is represented, and can not be directly connected with the accuracy of the model estimated LRT pollutants. This is emphasized in Section 3.3.

Lines 476-479: Speciation in SAPRC99 is unlikely to be the cause of model overestimates for O3. SAPRC99 underestimates VOC reactivity in chamber experiments, and the most recent updates to SAPRC are more reactive for urban than SAPRC99. For the rural CASTNet sites in this study, it

is more likely that overestimates of biogenic VOC in MEGAN and uncertainty in NOx emissions and fate contribute to the positive bias for O3.

This study does not cover any investigation on more recent updates in SAPRC, like SAPRC 07 or 11. In the text that we describe the amplified biases in STEM compared with the global models, here we just list some references that showed SAPRC99 produced much higher O_3 than other mechanisms which were used by certain HTAP2 global models. CASTNET O_3 is subject to the US urban O_3 pollution due to the regional-scale transport (See Huang et al., 2013b). In addition to the uncertainty from NO_x and BVOC emissions, we now extended the discussions to address a comment by Reviewer #3's: "Huang et al. (2017) showed that MEGAN's positive biases are in part due to the positively-biased temperature and radiation in WRF, and reducing ~2°C in WRF's temperature biases using a different land initialization approach led to ~20% decreases in MEGAN's isoprene emission estimates in September 2013 over some southeastern US regions…Quantifying the impacts of overestimated biogenic emissions and the biased weather fields that contributed to the biases in emissions on the modeled O_3 is still an ongoing work."

We also cited other regional model studies that attributed modeled biases to chemical mechanisms and emission biases: "Some existing studies also reported O_3 and NO_2 biases from other regional models in the eastern US, due to the chemical mechanism and biases in NO_x and biogenic VOC emissions (e.g., Canty et al., 2015)." And pointed out the need for future investigation on these from the AQMEII in the following sentence.

Lines 505-510: Note that larger-than-1 RERER values will be less likely to occur if the model results are analyzed for high O3 days. It is not informative to present results for low O3 days on which NO titration is more likely to occur because these days are not relevant to air quality attainment planning.

Again the monthly mean based analysis are shown to connect with the HTAP1 and other HTAP 2 modeling results (e.g., Surendran et al., AE, 2016 showed seasonal RERERs for HTAP2 in SAS) that are done on a monthly or seasonal basis. Although not directly relevant to the air quality management, these provide a broad characterization perspective (also considered in the 2016 EPA White Paper). We included a separate section (3.3) in the paper showing event based analysis.

Lines 516-519: "Comparing to the HTAP I modeling results, the magnitudes of R(O3, EUR, 20%) are smaller by a factor of 2-3, as a result of the substantial improvement in the European air quality over the past decades" The modeling for HTAP II is for 2010 versus 2001 for HTAP I, so any O3 reductions should reflect emissions reductions for 9 years, not for decades. Have European emissions been reduced by a factor of 2 to 3 from 2001 to 2010, or is it possible that other changes in the HTAP II modeling platform are the cause of this change?

In Section 2, we now have descriptions of "North Africa that are included in HTAP1's EUR domain. The impact of emissions over these regions on comparing the NAM R(O_3, EUR, 20%) values in HTAP1 and HTAP2 will be discussed in Section 3.2.1." And in the results section, following this commented sentence, we added: "… and also possibly due to the changes in the HTAP2 experiment setup from HTAP1 (e.g., EUR by HTAP1's definition includes regions in Russia/Belarussia/Ukraine, Middle East and North Africa that are excluded from the HTAP2's EUR domain)." We also believe this difference is in part due to the different HTAP1 and HTAP2 participating models and their configurations.

Lines 541-545: This text seems to inappropriately discount the significance of international transport and also the possible importance of differences among the global models. For interstate transport EPA uses 1% of the NAAQS as a significant contribution. Thus, differences among global model much less than 5% of the total model O3 can be very important, especially given that the values discussed in the text are based on a 20% emissions sensitivity and that results are reported as the monthly mean. While local emissions will have a larger contribution, it may not be true that local emissions control programs alone are the most effective way to attain the NAAQS, as the text seems to suggest.

The text in this paragraph has been modified to: "These monthly- and regional-mean R(O_3, EAS, 20%) values suggest that despite dilution along the great transport distance, the EAS anthropogenic sources still had distinguishable impact on the NAM surface O_3…. Also, similar to the findings from the HTAP1 studies, the large intermodel variability (as indicated in Table 4) in the estimates of intercontinental SR relationships indicates the uncertainties of these models in representing the key atmospheric processes which needs more investigations in the future. Overall, R(O_3, EAS, 20%) and its intermodel differences are much smaller than the biases of the modeled total O_3 in NAM. Other factors can contribute more significantly to the biases in the modeled total O_3, such as the stratospheric O_3 intrusion and the local O_3 formation, and assessing the impacts from these factors would be also helpful for understanding the uncertainties in the modeled O_3." Related sentences in Section 4 and the abstract were also revised.

Lines 562-571: It is surprising that the couple STEM/global model predicts large transport contributions than some global models and smaller transport contributions than other global models. The authors provide a list of factors that contribute to model uncertainty as a possible explanation, but it seems like these uncertainties (e.g., terrain, chemistry) should affect in similar ways each of the coupled STEM/global model simulations. More investigation is needed to explain why STEM sometimes shows higher or lower transport contributions compared to the global model.

This part now has been extended. The differences between regional and global models' results are due to the different terrain, met fields, transport and chemical production/loss between STEM and its boundary condition model. The differences of STEM/GC, STEM/CIFS and STEM/RAQMS pairs are different.

Lines 607-609: This is an important finding that should be highlighted in the conclusions and abstract.

This is now emphasized in both the abstract and the conclusions.

Lines 620-622: "Therefore, it is important for more HTAP2 participating models to save their outputs hourly in order to conveniently compute the policy-relevant metrics for the O3 sensitivities." I agree with this statement, and moreover, I do not think you can do a meaningful analysis of any models that do not save the hourly outputs (or 3-hour if that is the finest time resolution used), and I would recommend excluding them from this study.

Again the monthly mean based analysis are shown to connect with the HTAP1 and other HTAP 2 modeling results that are done on a monthly basis. Although not directly relevant to the air quality management, these provide a broad characterization perspective (also considered in the 2016 EPA White Paper). We included a separate section (3.3) in the paper showing event based analysis.

Lines 612-624 and Figure 9: It is obvious that day time O3 is greater than nighttime O3 at surface sites because O3 deposits to surfaces and is destroyed by chemical reactions at night. So the findings in this text that the maximum daily 8-hour average O3 is greater the 24-hour average O3 is self-evident. I suggest deleting this text. I also recommend focusing the analysis on maximum daily 8 hour averages, especially for the highest O3 days, and not showing results for monthly mean O3.

The original Figure 9 has been moved to the supplement, and the MDA8-based analyses have been extended. Figure 2c-d also show the diurnal cycles of the total O_3 and R (O_3, EAS, 20%) values.

Line 633: "R(MDA8, EAS, 20%) is smaller during the high O3 total days in all subregions." For GEOS-Chem the contribution appears to be the same on high O3 days compared to all days, and the results are very similar for RAQMS. It would be helpful to show more details for this analysis. Is this the mean O3 for all sites for days in which any monitor was > 70 ppb, or does it only include data for the monitor that was greater than 70 ppb? I suggest performing a more detailed analysis, e.g., show EAS contribution on each day for a few key sites that frequently have high O3, e.g., Great Basin and Canyonlands sites.

In the earlier version, regionally averaged (not only at CASTNET sites), and for each location/grid, only when the predicted total O_3 was over 70 ppbv. We now show averaged calculations and spatial plots at all CASTNET sites for all days and during the observed O_3 exceedances (Figures 11-12) and extended the discussions in text. These included Canyonlands and Great Basin. Canyonlands NP is also one of the sites that experienced O_3 exceedances on 9 May (Section 3.3).

Line 655: "We found that the underestimated free tropospheric O3 from the STEM simulations that used any single free-running chemical boundary conditions contributed to the underestimated STEM surface O3 in the high elevation mountain states." Need to edit and clarify meaning of the above sentence. Was this because the global models underestimated stratospheric O3 or international transport?

It could be a result of the underestimation of both. The possible uncertainty in LRT of Asian pollution was determined with the help of evaluation at the source side. We also added a sentence in this paragraph about the limitation of models representing the stratospheric intrusion: "As the enhancement of O_3 due to the assimilation is much larger than the O_3 sensitivities to the EAS anthropogenic emissions, the assimilation mainly improved the contributions from other sources, such as the stratospheric O_3."

References (those not cited in the text but in this response file)

Choi, Y., J. Kim, A. Eldering, G. Osterman, Y. L. Yung, Y. Gu, and K. N. Liou (2009), Lightning and anthropogenic NO_x sources over the United States and the western North Atlantic Ocean: Impact on OLR and radiative effects, Geophys. Res. Lett., 36, L17806, doi:10.1029/2009GL039381.

Surendran, D. E., S. D. Ghude, G. Beig, C. Jena, D.M. Chate (2016), Quantifying the sectoral contribution of pollution transport from South Asia during summer and winter monsoon seasons in support of HTAP-2 experiment, Atmos. Environ., 145, 60-71, doi:10.1016/j.atmosenv.2016.09.011.

Wolfe, G. M., T. F. Hanisco, H. L. Arkinson, T. P. Bui, J. D. Crounse, J. Dean-Day, A. Goldstein, A. Guenther, S. R. Hall, G. Huey, et al. (2015), Quantifying sources and sinks of reactive gases in the lower atmosphere using airborne flux observations, Geophys. Res. Lett., 42, 8231–8240, doi:10.1002/2015GL065839.