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."

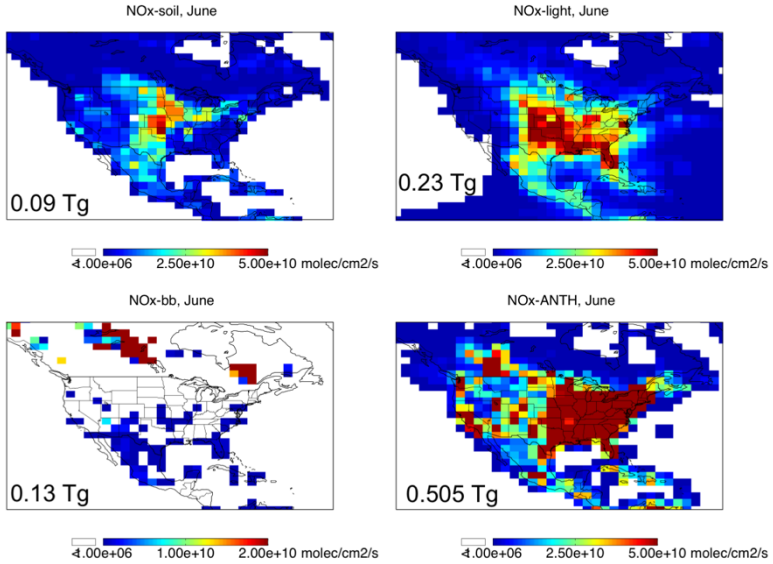
Line 143: "the necessity of evaluating the extra-regional source impacts on event scale [have] has been emphasized" This is a key point – check to see if this addressed in discussion and conclusions.

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

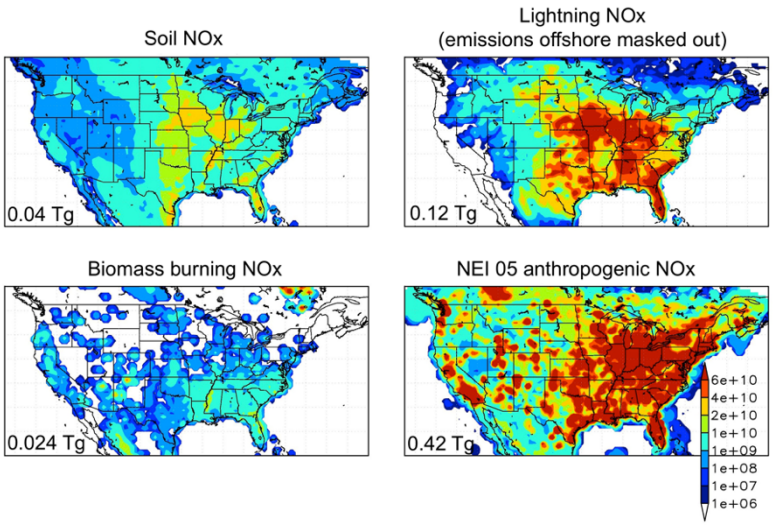
Lines 214-216: Biogenic emissions of VOC are larger than anthropogenic VOC globally, and biogenic and geogenic emissions of NOx, SO2, CO and CH4 are also large and can have a substantial impact on model results. It would have been best to harmonize the natural emissions in addition to anthropogenic emissions, 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.

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.

Response to Reviewer #3's comments

We thank the careful review by Reviewer #3. Please see below our response (in blue) to the general and specific comments (in black).

General Comments

This manuscript presents the first HTAP Phase II findings, expanding on HTAP Phase I by incorporating regional models to estimate the impact of international anthropogenic emissions on U.S. surface ozone. The authors use boundary conditions from three different global models to drive the regional STEM model, and compare the sensitivities of surface ozone in North America to international anthropogenic emissions with those determined from 8 global models. They further compare with an adjoint version of one model, use boundary conditions from a model that assimilated satellite ozone products, and conduct a case study using multiple satellite and ground-based products. This is a major undertaking, as noted by another reviewer. I agree with the other reviewers, however, that the paper suffers from some shortcomings. Several of the issues I was planning to cover were discussed at length in the earlier reviews, so I focus below on additional points. I'd like to see the abstract/conclusions clarify and quantify (e.g., within 10%, 30%, factors of 2?) the conclusions regarding how different the global and regional model estimates are, and how much the RER sensitivity estimates have changed from those reported in the 2010 HTAP report. I agree with Dr. Tonnesen that more emphasis on episodic events would enhance the policy relevance of this work. Throughout the text, more quantitative and specific language should be used wherever possible, and the paper should be edited carefully for clarity (e.g., incomplete sentence L768). The introduction is quite long and could state earlier on what the point of this study is to provide a context before going into all the details of past work.

Quantitative language is now used in places as you suggested, especially in the abstract and the conclusions. The R values were compared with the HTAP1 results in the 2010 report, while the RERER calculation is a new element in HTAP2. Taking Dr. Tonnesen's suggestion, we added a summer event case study, drawing some similar conclusions to the 9 May event. We also show averaged calculations and spatial plots at all CASTNET sites for all days and during the observed O_3 exceedances (Figures 11-12).

Manuscript has been carefully and extensively edited. The sentence starting at L768 now reads as: "As emissions from various source sectors can differ by emitted altitudes and temporal profiles, efforts should also be placed to have the models timely update the height and temporal profiles of the emissions from various sectors." The last paragraph of the introduction section was substantially modified, with the specific goals of this study stated first and some details of the methods moved to later sections.

Specific Comments L42-45. Elaborate on what this means for drawing conclusions regarding the role of hemispheric transport of air pollution.

We found that the differences between STEM surface O_3 sensitivities and its corresponding boundary condition model's are often smaller than those among its boundary condition models. We also reported the differences between boundary models and all global models. We agree that these are key findings of this paper, indicating that accurately attributing pollution in the global model(s), which still appeared to be difficult, is a critical first step for any follow-on estimates

based on regional models using the boundary conditions downscaled from these global models. We have rewritten this (and later) sentences.

L48 'Tagged tracer approach' is mentioned here and elsewhere (e.g. L564); a brief explanation is needed as approaches can involve tagging ozone itself or tagging precursors.
This is a good point. The Asian O_3 in Brown-Steiner and Hess (2011) means any O_3 created as a result of anthropogenic+biofuel NO_x emissions (with no interannual variability) over the East Asia. They should be compared with EAS NO_x emission perturbation runs. However, we here only used sensitivities to EAS NO_x emission perturbation from GEOS-Chem, so the direct comparisons in the abstract and multiple place in the text were removed. The seasonality based on tagging and 20% NO_x emission perturbation was compared in Section 3.2.1 instead.

I'm not convinced that this study cleanly isolated the role of rising East Asian anthropogenic emissions; see also RC2 comments.
Same as the response to RC2's comments:
The comparisons of HTAP1 and HTAP2 findings over larger spatial/temporal scales in this study 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

L51 Are the adjoint sensitivities compared to all the global models or just the forward version of GEOS-Chem? Is this the same version as used to provide boundary conditions? (see also L591)
Just GEOS-Chem's. The CU and SNU GEOS-Chem are different.

L54-56 Try to quantify this statement: is it off by 20%? Factor of 2?
Done.

L57-59 This appears to be a general statement rather than a conclusion drawn from this work and thus does not seem appropriate to include in the abstract.
More conclusive language is now used to describe the findings from satellite data related work in this paper.

L96 The first paper to show this was Jacob et al., GRL, 1999: http://onlinelibrary.wiley.com/doi/10.1029/1999GL900450/abstract
Cited.

L148. Region-dependent, but also time-dependent?
Added "time-".

L220-227 Seems relevant to provide BVOC emissions over Asia and North America. How much do North American anthropogenic emissions contribute to global totals?
North American anthropogenic emissions contribute to global totals can be calculated by numbers in Table S1. A sentence is added to Section 2.1: "In 2008, NAM NO_x, NMVOC and CO contributed to 18.0%, 11.7% and 11.9% of the global total, respectively, and in 2010, these contributions became 15.8%, 10.5% and 10.2%.".

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."

L233 References could be included in Table 1
Done. Related text and Table 1 caption was modified accordingly.

L238 Why are boxes shown in Figure 1 if the regions are actually following the political boundaries as indicated in L258?
The boxes were used to highlight the three focused source regions (EAS, SAS, EUR) rather than defining the boundaries of these regions, as mentioned in the figure caption. To avoid the confusion, we instead highlighted these three regions by underlining the region names in the map.

L276. Given that Lin et al. 2012 estimated Asian ozone pollution transport to the western U.S. using a global model about this resolution, a case needs to be made for why it's appropriate to use a regional model (e.g., allows testing of multiple boundary conditions, and regulatory applications would presumably run at finer scales).
Lin et al. (2012a) used a different model (parameterizations are different) with different configurations (e.g., the emission input). They mainly focused on the western US, and the impact of data assimilation on the modeled O_3 was not addressed in that study. As included in the discussions, all R values during the exceptional events are smaller than 1/5 of their reported sensitivities, due to the differences in model parameterizations and configurations. Some related discussions can be found in Section 3.3.

In terms of the use of regional models, we agree with your suggestion that it allowed us to test the multiple boundary conditions. And, same as our response to Reviewer #2's comment: In Section 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, 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 its resolution and other configurations, parameterization) to accurately estimate the 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.

We also make the readers be aware 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…."

L283. This may be true for the Asian pollution transport, but Lin et al. 2015 indicate that 2010 isn't a particularly high year for stratospheric intrusions reaching surface air over the WUS. http://www.nature.com/articles/ncomms8105 See their figure 2.
This part has been modified to focus on the interannual variability of LRT of Asian pollution.

L287-88 Is this just reflecting the warming trend over the 81-00 period? Temperatures and ozone production were even higher in 2011 and 2012 in the eastern US.
Yes. As it's based on "the climatology from the NCEP/NCAR reanalysis data for the 1981-2010" described earlier in this paragraph. This paper does not cover the periods after 2010.

L315 How was this downscaling done?
Standard downscaling approach: spatial/temporal interpolation and species mapping.

L442 perhaps needs a reference for the HTAP1 work unless this was done as part of this study?
Added Fiore et al. (2009).

L445-447. It's not clear what the take-away point is here. Are the models underestimating Asian pollution influence or can we not tell because it could be regional transport? Presumably even though the data assimilation fixes this problem, it does not help us to distinguish between these possible sources of error?
This sentence just lists the possible sources of error, including both trans-boundary (see case study for details) and regional transport, but it does not distinguish/quantify the impact from each factor.

L472. How did this study determine that the bias is likely due to overestimated anthropogenic NOx emissions? May doesn't look like it has a clear bias whereas July does. How do we know this is associated with anthropogenic sources rather than seasonally varying sources like soil NOx for example? Are there seasonal variations in the anthropogenic NOx emissions?
Both anthropogenic and non-anthropogenic emissions are time-varying. Anthropogenic emissions differ by month (Section 2.1) while many non-anthropogenic emissions are weather dependent and display stronger temporal varibility. Overall anthropogenic NO_x emissions contribute most to the total NO_x emissions, but the uncertainty can definitely be due to those from other emission sources. We now added to the SI the natural emissions from GEOS-Chem and STEM in June 2010, and mentioned about the possible overprediction in soil/lightning NO_x in the central/eastern US near L472: "Larger OMI-model disagreement was found over the central/eastern US during June 2010, likely also due to the uncertainty in GEOS-Chem's soil or lightning NO_x emissions, which appear to be high over these regions (Figure S1)".

L481-483. Doesn't this interpretation depend on where the photochemical regime is at in terms of ozone production with respect to NOx emissions?
We added "Under different chemical regimes," before "this statement would also rely on the quality of other O_3 precursors in the HTAP2 emission inventory..".

L520 An estimate of how large these biases are and how much bias they introduce into ozone would be useful here.
The biases are time- and region- dependent and in part depend on the quality of the WRF inputs. We added the findings from Huang et al. (2017) on the Sep 2013 conditions for MO and TX 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.

L541-544. There seems to be model disagreement near the Canadian border, with Oslo for example suggesting high cross-border influence but CHASER suggesting much less. L544-547. Is Oslo also higher resolution as it looks similar to EMEP in terms of higher influence.
OsloCTM3's horizontal resolution is 2.8°×2.8° (Table 1a), but we noticed that the number of its vertical layers, which affects the export and import of pollution, are larger than the rest of the models'. The number of vertical layers for each model are now added to Table 1. We added: "Although on a coarse horizontal resolution of 2.8°, OsloCTM3 suggests stronger extra-regional source influences on the northwestern US and the US-Canada border regions than the other models. Its largest number of vertical layers among all global models might be a cause."

L585-588 Where is this shown?
This is a general statement pointing out the other key sources.

L612-616. Can you provide estimates of how the ozone lifetimes in the boundary layer differ in the different simulations?
This is a good suggestion that we did not prepare for HTAP2 and is in need for future analysis.

L627-628. There seems to be an assumption that LRT is obvious from satellite data. This isn't the case for ozone. How will LRT be convincingly separated out from other ozone sources?

All observations, not only the satellite observations, represent the total O_3. The use of satellite O_3 and CO can distinguish anthropogenic/biomass burning sources from the stratospheric intrusions, and additional tools and data will also be helpful. However, this sentence is to say that the broader coverage provided by the future satellites (than the CASTNET network) would better help capture polluted events. As Dr. Lin also pointed out in their 2015 GRL paper, the sampling strategy does affect the calculated pollution trends and source attribution, and in the paper we compared sensitivities in all grids v.s. only at CASTNET sites.

L658-659. Did all models capture the same events in terms of their timing and approximate regional location?
Qualitatively similar. We added "(based on three boundary condition models separately and averagely)". See revised Figure 2b (the thin lines show individual models' EAS sensitivity for the western US) and the case studies for the detailed comparisons (e.g., Figures 15/18).

L661. It would be more convincing to show this as a monthly mean diurnal cycle rather than rely on Figure 2a.
Time series in Figure 2b (previously Figure 2a) shows the 3-6 LRT events during May-June 2010. Period-mean diurnal cycles are now also shown in Figures 2c-d for total O_3 and the EAS sensitivities, respectively.

L679-681. I didn't follow this point.
STEM base simulations overall substantially overpredicted the total O_3 in non-western US regions based on our evaluation at the CASTNET sites, as described in the previous sections. So the R(MDA8, EAS, 20%) calculated during the days of O_3 exceedances (based on the STEM-estimated total O_3 in all model grids) can actually represent the sensitivities during some days when total O_3 actually did not exceed 70 ppbv. We now also added that some of the exceedances in the western US were not correctly captured which also affected conclusions from this figure.

L741-742. Be more specific here.
Quantitative results are summarized here and in the abstract, which also addressed your general comment.

L744-747. Is there a relationship between the bias and the Asian transport events?
The biases in modeled total O_3 are attributed to those in the modeled LRT Asian pollution as well as other factors. But the model that predicted the higher O_3 does not always gave higher estimates of the EAS contribution, as shown in the case study.

L747-749. It's not clear how better quantifying stratospheric o3 intrusion helps reduce North American pollution levels and model uncertainties. This statement also implies that stratospheric intrusions are as important as local ozone formation.
This paragraph has been rewritten to suggest impact from bottom-up emission input and future work on attributing the intermodel differences and model biases.

L750. How frequent are these episodic sensitivities to East Asian emissions? Are they occurring when measured ozone is highest?

This paragraph has been rewritten based on the additional analyses we performed for high O_3 days: "The STEM O_3 sensitivities to the East Asian anthropogenic emissions (based on three boundary condition models separately and averagely) were strong during 3-6 episodes in May-June 2010, following similar diurnal cycles as the total O_3. Stronger-than-normal East Asian anthropogenic pollution impacts were estimated during O_3 exceedances in the western US, especially over the high terrain rural/remote areas; in contrast, non-local pollution impacts were less strong during O_3 exceedances in other US regions."

L800-801. These suggestions seem to neglect the important caveat that these approaches assume that model transport is perfect.

This depends on what kind of model(s) to be used. For online models, weather fields may be modified together with the chemical fields; For offline models, you are right, but these suggested methods still incrementally improve the source attribution and should be encouraged.

Response to Dr. Lin

We thank the additional references suggested by Dr. Lin and have made changes to the manuscript accordingly. Please see below our responses in blue.

1. Lines 93-100, Page 3: Regarding Asian influence on US ozone trends, please consider citing the following papers and discuss their findings:

Lin, M., L.W. Horowitz, O.R. Cooper, D. Tarasick, S. Conley, L.T. Iraci, B. Johnson, T. Leblanc, I. Petropavlovskikh, E.L. Yates (2015): Revisiting the evidence of increasing springtime ozone mixing ratios in the free troposphere over western North America, Geophysical Research Letter, 42, doi:10.1002/2015GL065311

Lin, M., W. Horowitz, R. Payton, A.M. Fiore, G. Tonnesen. US surface ozone trends and extremes over 1980-2014: Quantifying the roles of rising Asian emissions, domestic controls, wildfires, and climate. Atmos. Chem. Phys. Discuss., doi:10.5194/acp-2016- 1093, 2016

You cited Cooper et al. (2010, Nature). But Lin et al. (2015 GRL) investigated the representativeness of ozone trends derived from sparse measurements reported by Cooper et al. They found that sampling biases can substantially influence calculated ozone trends.

Both papers are now cited in this paragraph. We do agree that the observation sampling methods affect the calculated pollution trends and SR relationships, and related discussions have been included in several places in Section 3, e.g., comparing R values averaged over all grids and only sampled at the CASTNET sites.

2. The multi-model results presented in this article are based on the spring of 2010 following strong El Nino conditions. I think it would be useful to the readers if you can discuss the representativeness of your results on inter-annual context. There are studies showing that long-transport transport of Asian pollution is stronger during El Nino springs due to the eastward extension and equator-ward shift of the subtropical jet stream (e.g., Lin et al., 2014, Nature Geoscience).

We agree that discussing the results on inter-annual context is important. Interannual differences of LRT of Asian pollution due to the impact of atmospheric circulation v.s. anthropogenic emission changes are briefly discussed in Sections 2 and 3.2, and are highlighted in the abstract and conclusions. A sentence has been added to Section 2.1 mentioning the findings in Lin et al. (2014): "This is consistent with the findings by Lin et al. (2014) that the El Niño conditions during the 09/10 winter strengthened the trans-Pacific transport of Asian pollution in spring 2010."

Impact of Intercontinental Pollution Transport on North American Ozone Air Pollution: An HTAP Phase 2 Multi-model Study

Min Huang[1,2], Gregory R. Carmichael[3], R. Bradley Pierce[4], Duseong S. Jo[5], Rokjin J. Park[5], Johannes Flemming[6], Louisa K. Emmons[7], Kevin W. Bowman[8], Daven K. Henze[9], Yanko Davila[9], Kengo Sudo[10], Jan Eiof Jonson[11], Marianne Tronstad Lund[12], Greet Janssens-Maenhout[13], Frank J. Dentener[13], Terry J. Keating[14], Hilke Oetjen[8,*], Vivienne H. Payne[8]

[1]George Mason University, Fairfax, VA, USA
[2]University of Maryland, College Park, MD, USA
[3]University of Iowa, Iowa City, IA, USA
[4]NOAA National Environmental Satellite, Data, and Information Service, Madison, WI, USA
[5]Seoul National University, Seoul, Korea
[6]European Center for Medium range Weather Forecasting, Reading, UK
[7]National Center for Atmospheric Research, Boulder, CO, USA
[8]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA
[9]University of Colorado-Boulder, Boulder, CO, USA
[10]Nagoya University, Furocho, Chigusa-ku, Nagoya, Japan
[11]Norwegian Meteorological Institute, Oslo, Norway
[12]Center for International Climate and Environmental Research, Oslo, Norway
[13]European Commission, Joint Research Center, Ispra, Italy
[14]US Environmental Protection Agency, Washington, DC, USA
[*]Now at: University of Leicester, Leicester, UK

Correspondence to: Min Huang (mhuang10@gmu.edu)

Deleted: II

Deleted: Research Triangle Park, NC

Formatted: Suppress line numbers

Abstract

The recent update on the US National Ambient Air Quality Standards of the ground-level ozone (O_3) can benefit from a better understanding of its source contributions in different US regions during recent years. In the Hemispheric Transport of Air Pollution experiment Phase 1 (HTAP1), various global models were used to determine the O_3 source-receptor relationships among three continents in the Northern Hemisphere in 2001. In support of the HTAP Phase 2 (HTAP2) experiment that studies more recent years and involves higher-resolution global models and regional models' participation, we conduct a number of regional scale Sulfur Transport and dEposition Model (STEM) air quality base and sensitivity simulations over North America during May-June 2010. STEM's top and lateral chemical boundary conditions were downscaled from three global chemical transport models' (i.e., GEOS-Chem, RAQMS, and ECMWF C-IFS) base and sensitivity simulations in which the East Asian (EAS) anthropogenic emissions were reduced by 20%. The mean differences between STEM surface O_3 sensitivities to the emission changes and its corresponding boundary condition model's are smaller than those among its boundary condition models, in terms of the regional/period mean (<10%) and the spatial distributions. An additional STEM simulation was performed in which the boundary conditions were downscaled from a RAQMS simulation without EAS anthropogenic emissions. The scalability of O_3 sensitivities to the size of the emission perturbation is spatially varying, and the full source contribution obtained by linearly scaling the North American mean O_3 sensitivities to a 20% reduction in the EAS anthropogenic emissions may be underestimated by at least 10%.

The three boundary condition models' mean O_3 sensitivities are ~8% (May-June 2010)/~11% (2010 annual) lower than those estimated by multiple global models, and the multi-model ensemble estimates are higher than the HTAP1 reported 2001 conditions, due to the growing EAS anthropogenic emissions, the interannual variability in atmospheric circulation (i.e., stronger trans-Pacific transport in spring 2010 following an El Niño event), and the different experiment designs of HTAP1 and HTAP2. GEOS-Chem sensitivities indicate that the EAS anthropogenic NO_x emissions matter more than the other EAS O_3 precursors to the North American O_3, qualitatively consistent with previous adjoint sensitivity calculations.

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. Satellite O_3 (TES, JPL-IASI, and AIRS) and carbon monoxide (TES and AIRS) products, along with surface measurements and model calculations, show that during certain episodes stratospheric O_3 intrusions and the transported EAS pollution influenced O_3 in the western and the eastern US differently. Free-running global models underpredicted the transported background O_3 during these episodes, posing difficulties for STEM to accurately simulate the surface O_3 and its source contribution. Although we effectively improved the modeled O_3 by incorporating satellite O_3 (OMI and MLS) and evaluated the quality of the HTAP2 emission inventory with the KNMI OMI nitrogen dioxide, using observations to evaluate and improve O_3 source attribution still remains to be further explored.

Deleted: The STEM

Deleted: (

Deleted: %). We perform analyses not only on large spatial/temporal scales relative to the HTAP1, but also on subcontinental- and event-scale that are more relevant to the US air quality management. The

Deleted: (including the 24h mean and the US policy-relevant maximum daily 8h average (MDA8) metric averaged throughout the study period and during a selected pollution transport event)

Deleted: often

Deleted: . The STEM sensitivities are also compared with the mean sensitivities estimated by multi- global models, which are higher than the HTAP1 reported 2001 conditions, as well as the 2001-2005 conditions studied using the tagged tracer approach. This indicates the increasing impacts

Deleted: East Asian anthropogenic pollution on North America during 2001-2010. The GEOS-Chem sensitivities indicate that the East Asian anthropogenic NO_x emissions matter more than the other East Asian O_3 precursors to the North American O_3, qualitatively consistent with previous adjoint sensitivity calculations

Deleted: global

Deleted: East Asian

Deleted: , to assess the

Deleted: . The scalability

Deleted: regional

Deleted: y

Deleted: the

Deleted: East Asian

Deleted: .

Deleted: Satellite NO_2 (KNMI OMI) and O_3 (TES, JPL-IASI, OMI, MLS, and AIRS) products help detect pollution episodes, quantify or/and reduce the uncertainties in the bottom-up NO_x emissions and the model transported background O_3. Based on model calculations and satellite/surface observations during a selected event, we show the different influences from stratospheric O_3 intrusions along with the transported East Asian pollution on O_3 in the western and the eastern US. Future directions of using satellite data in air quality research are also suggested.

Formatted: Suppress line numbers

1. Introduction

Tropospheric ozone (O_3), a short-lived trace gas with a lifetime ranging from hours in the boundary layer to weeks in the free troposphere, affects tropospheric chemistry, harms human and ecosystem health, and induces climate change on local, regional and global scales (Jerrett et al., 2009; Smith et al., 2009; Anenberg et al., 2010; Mauzerall and Wang, 2001; Avnery et al., 2011a, b; Shindell et al., 2009, 2013; Bowman and Henze, 2012; Stevenson et al., 2006, 2013; Monks et al., 2015). It has been recognized that the uneven distributions of tropospheric O_3 can be attributed to the stratosphere as well as local, regional and distant emission sources, through complicated processes that occur on synoptic, meso- and micro-scales (Task Force on Hemispheric Transport of Air Pollution (HTAP), 2010; National Research Council (NRC), 2009; Maas and Grennfelt, 2016). The mitigation of O_3's climate and health impacts would benefit from efforts to control the emissions of its precursors from the various emission sources (United Nations Environment Programme (UNEP) and World Meteorological Organization (WMO), 2011), such as nitrogen oxides (NO_x), carbon monoxide (CO), methane (CH_4), and non-methane volatile organic compounds (NMVOCs).

Ground-level O_3 is one of the six criteria air pollutants regulated by the US Environmental Protection Agency (EPA), and the US National Ambient Air Quality Standards (NAAQS) has recently been lowered to 70 ppbv to better protect Americans' health and the environment. 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). This includes assessing the impacts of various components of the background O_3, such as stratospheric O_3, local natural sources such as biogenic, lightning and wildfire emissions, as well as the long-range transport (LRT) of pollution. The impact of the trans-Pacific pollution transport on US air quality has been evaluated in numerous studies over the past decades (e.g., Fiore et al., 2009; Reidmiller et al., 2009; Zhang et al., 2008, 2009; Huang et al., 2010, 2013a; Lin et al., 2012a, 2015, 2016; US EPA, 2016a). It has been found that the increasing trends of pollution in the upwind continents, especially the populated East Asia (e.g., Zhang et al., 2014; Susaya et al., 2013; Wang et al., 2012), may partially offset the US air quality improvements in recent decades due to the regional and local emission controls (e.g., Jacob et al., 1999; Verstraeten et al., 2015; Ambrose et al., 2011; Wigder et al., 2013; Cooper et al., 2010; Parrish et al., 2009, 2012; Gratz et al., 2014). A better understanding of the processes that determine the O_3 pollution levels, as well as an improved capability of attributing the air pollution to nearby or distant sources is needed to assist with designing and implementing effective local emission control strategies to comply with the tighter air quality standards.

Chemical transport models are often used to reproduce and attribute the observed O_3 levels, including assessing the impacts of the internationally transported O_3 on the US air quality. In the HTAP modeling experiment Phase 1 (HTAP1), various global models with horizontal resolutions ranging from 1°×1° to 5°×5°, only around half of which are finer than 3°×3°, were used to determine the O_3 source-receptor (SR) relationships among three continents in the Northern Hemisphere in 2001 (Chapter 4 in HTAP, 2010). The global model based SR relationships in HTAP1 determined using the emission perturbation approach (i.e., calculating the changes of O_3

Deleted: I

at the receptor regions in response to a 20% reduction in the emission inputs in a given source region) were reported as either monthly 24h mean values or policy-relevant metrics such as the maximum daily 8h average (MDA8) for the US (e.g., Fiore et al., 2009; Reidmiller et al., 2009). Large intermodel diversity was found in the simulated total O_3 and the intercontinentally transported pollution for the chosen SR pairs in the northern midlatitudes, indicating the challenges with model simulations to accurately represent the key atmospheric processes. Multi-model mean results were the foci of in these studies with the assumption that this approach can reduce the uncertainty from the single model estimates for monthly or seasonal means. "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). Although the multi-model approach can help identify some of the weaknesses of the individual models and may produce more reliable estimates, it is necessary to well understand the uncertainties inherent in using the same set of anthropogenic emissions in all these model simulations. Satellite observations over the regions with limited in-situ measurements such as the East Asia can be particularly helpful for quantifying such uncertainties.

The 20% emission perturbation in the HTAP1 modeling experiment was chosen to produce a sizeable (i.e., larger than numerical noise) and realistic impact, but small enough in the assumed near-linear atmospheric chemistry regime. The scalability of the modeled O_3 sensitivities to the size of the emission perturbations has been assessed on continental scale (Wu et al., 2009; Fiore et al., 2009; HTAP, 2010; Wild et al., 2012; Emmons et al., 2012). The receptor O_3 responses to the source-region emission perturbations are found to be fairly linear within ~50% of the perturbations. However, due to the chemical non-linearity, the full source contribution obtained by linearly scaling the receptor regional mean O_3 sensitivity to the 20% reduction in the source region emissions may be underestimated, and the scalability depended on seasons and the perturbed emission species. Huang et al. (2013b) investigated the scalability of the O_3 sensitivity between the southern California-US intermountain west SR pair for May 2010, in which study the southern California anthropogenic emissions were perturbed by multiple amounts of +50%, -50%, -100%. They reported that the scalability of the O_3 sensitivities changed with the distance from the source regions. Further analyses on the scalability of these modeled O_3 sensitivities during recent years especially for the East Asia-NAM SR pair, as well as their spatial variability, are still needed. Furthermore, results generated using the emission perturbation approach need to be compared with those based on the other methods (e.g., tagged tracers, adjoint sensitivity).

Previous studies have demonstrated the advantages of high resolution chemical transport modeling for understanding SR relationships (e.g., Lin et al., 2010 for Europe and the East Asia; Lin et al., 2012a; Huang et al., 2010, 2013a for Asia and NAM). Using observations (satellite, sondes, aircraft) along with single model simulations, a few studies have reported that the US O_3 sensitivities to extra-regional sources is time- and region-dependent (e.g., Lin et al., 2012a, b; Langford et al., 2011; Ott et al., 2016), and therefore the necessity of evaluating the extra-regional source impacts on event scale has been emphasized in these studies as well as in US EPA (2016a, b). The HTAP Phase 2 (HTAP2) multi-model experiment, initiated in 2012, is designed to advance the understanding of the impact of intercontinental pollution transport during more recent years (i.e., 2008-2010) involving a number of global and regional models' participation (Galmarini et al., 2017; Koffi et al., 2016). The regional models are anticipated to help connect the analyses over global and regional scales and enable discussions on small spatial (e.g., subcontinental) and

Deleted: 6

temporal scales (i.e., event based analyses). The use of satellite products for identifying the transport events as well as for quantitative model evaluation is also encouraged in the work plan. The HTAP2 modeling experiment was sequentially conducted in two steps. First, similar to the HTAP1 experiment, a group of global models with different resolutions conducted base and emission perturbation sensitivity simulations to determine the pollutants' SR relationships. All models in their base simulations used the same set of harmonized sector-based global anthropogenic emissions developed specifically for the HTAP2 modeling experiment (Janssens-Maenhout et al., 2015). Most of these global models recorded only key chemical species from their base and sensitivity simulations in varied temporal frequencies. Several global models saved the three-dimensional (3D) chemical fields of more species with a 3- or 6-hour interval, which are suitable for being used as regional models' chemical boundary conditions. In the second step, regional models conducted base and sensitivity simulations to analyze the pollutants' SR relationships in greater detail. The regional model simulations used the same set of anthropogenic emissions as the global models within their simulation domains, and the chemical boundary conditions in these regional simulations were downscaled from the base and sensitivity simulations from the selected boundary condition model outputs. 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.

This study aims to address: 1) the differences in O_3 sensitivities generated from the HTAP2 and HTAP1 experiments to help address how the LRT impacts on NAM changed through time; 2) 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; 3) the usefulness of satellite observations for better understanding the sources of uncertainties in the modeled total O_3 (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. We performed a number of regional scale STEM (Sulfur Transport and dEposition Model) base and sensitivity simulations over the NAM during May-June 2010, during which period strong trans-Pacific pollution transport were shown to episodically impact the US (Lin et al., 2012a). 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) base and sensitivity simulations in which the East Asian anthropogenic emissions were reduced. The STEM surface O_3 sensitivities over the NAM region based on different boundary condition models were inter-compared, in terms of the regional averages and the spatial patterns on monthly basis and during a selected event identified by satellite O_3 and CO products. These were also compared with the sensitivities estimated by their corresponding boundary condition models as well as all HTAP2 participating global models and the results from HTAP1.

2. Methods

2.1. Anthropogenic emission inputs

Identical anthropogenic emissions were used in all global and regional chemical transport models' base and sensitivity simulations. This monthly-varying harmonized sectoral (i.e., power, industry, transportation, residential, shipping, aircraft, agriculture) emission inventory was

Deleted: In this study, we

Moved (insertion) [1]

Moved up [1]: 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.

Deleted: . An additional regional simulation was performed in which the STEM boundary conditions were downscaled from one global model simulation without the East Asian anthropogenic emissions, and the nonlinear relationship between the O_3 sensitivity and the size of the emission perturbation is discussed. In the discussion, we emphasize: 1) the differences in O_3 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 O_3 (e.g.,

provided on a gridded 0.1°×0.1° resolution for the years of 2008 and 2010, by compiling the officially reported emissions at the national scale (Janssens-Maenhout et al., 2015; http://edgar.jrc.ec.europa.eu/htap_v2). The temporal profiles for developing the monthly-varying emissions differ by region and sector. The amount of emissions of key O_3 precursors (CO, NO_x, NMVOCs) from both years are summarized in Table S1 for the four major emissions sectors, over the NAM (US+Canada, based on data from the US EPA and the Environmental Canada, which shows lower emissions from the previous years as also discussed in Pouliot et al., 2015), 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), and for over the world. For all of these species, global total emissions in 2008 and 2010 are similar. The NO_x, NMVOC, and CO emissions decreased from 2008 to 2010 over the NAM by 10.7%, 9.4%, and 15.7%, respectively. In 2008, NAM NO_x, NMVOC and CO contributed to 18.0%, 11.7% and 11.9% of the global total, respectively, and in 2010, these contributions became 15.8%, 10.5% and 10.2%. For 2010, the transportation sector contributed more than the other sectors to NAM anthropogenic NO_x and CO emissions; industrial sector contributed more than the other sectors to NMVOCs emissions. Over East Asian countries, these emissions are ~2-5 times higher than the US emissions, and the NO_x, NMVOC and CO emissions increased over Asia by 7.3%, 7.2% and 1.0%, with the dominant emission sectors in 2010 of transportation, industry, and residential, respectively. For both years, the emissions over the MICS-Asia regions contribute to over 40% of the global emissions. For these key O_3 precursors, the East Asian countries contribute to 45% (NMVOCs)-70% (NO_x) of the emissions in the MICS-Asia domain in both years, and the south Asian countries contribute to ~22% (NO_x)-34% (NMVOCs) of the MICS-Asia emissions. The uncertainty of the emission estimates differs by emission sector and species: i.e., the emissions from large-scale combustion sources (e.g., NO_x and CO from power and industry sectors) are less uncertain than those from small-scale and scattered sources (e.g., CO and NMVOCs from transportation and residential sources). 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.

2.2. Region definitions for the SR study and the model base and sensitivity simulations

2.2.1. Base and 20% emission perturbation simulations from global and regional models

The HTAP2 simulations from eight global models, used in this study, are listed in Table 1a, including the relevant references. Horizontal and vertical resolutions of these models range from finer than 1° to coarser than 2.5°, and from 20 to 60 layers, respectively. Overall these resolutions are higher than the HTAP1 participating models'. Figure 1 defines the source regions used in the HTAP2 SR relationship study and we will focus in this study on assessing the East Asia (EAS), S Asia (SAS), Europe (EUR), and non-NAM anthropogenic source (interchangeable in this paper with "(all) foreign") impacts on the NAM O_3 levels in 2010. Specifically, each model performed a base simulation and a number of sensitivity simulations in which the original HTAP2 anthropogenic emissions for all species and sectors in a defined source region were perturbed by a certain amount (referring to 20% as in most cases) and these cases are defined in Table 1a-b as **source region**ALL(**perturbation**), where "ALL" refers to "all species and sectors", consistent with HTAP1 and HTAP2's naming convention. The O_3 differences R(O_3, **source region**, **perturbation**) over the NAM were then calculated between each model's base and sensitivity simulations:

Deleted: 5

Deleted: . As this paper focuses

Deleted: impact of anthropogenic emissions, we do not introduce this information

Formatted: Indent: First line: 0.5"

Deleted: .

Deleted: overall

Deleted: Relevant references for the RAQMS model and the SNU GEOS-Chem are Pierce et al. (2007, 2009) and Park et al. (2004) (with additional descriptions on its HTAP simulation configurations at: http://iek8wikis.iek.fz-juelich.de/HTAPWiki/WP2.3?action=AttachFile&do=view&target=_README_GEOS-Chem.pdf), respectively. The descriptions of the remaining models can be found in published HTAP2 works such as in Stjern et al. (2016).

Deleted: , unless specified differently

Deleted: 1

Deleted: (

Deleted:).

$R(O_3, EAS, 20\%) = BASE\ O_3 - EASALL(-20\%)\ O_3$ (1a)

$R(O_3, SAS, 20\%) = BASE\ O_3 - SASALL(-20\%)\ O_3$ (1b)

$R(O_3, EUR, 20\%) = BASE\ O_3 - EURALL(-20\%)\ O_3$ (1c)

$R(O_3, non\text{-}NAM, 20\%) = NAMALL\ O_3 - GLOALL(-20\%)\ O_3$ (1d)

The monthly-mean R(O_3, **source region**, 20%) values were averaged over the NAM region for the analysis and compared with the findings in the HTAP1 study (e.g., Fiore et al., 2009). It is worth mentioning that the rectangular source regions defined in HTAP1 were modified in HTAP2 to align with the geo-political borders. For EAS and SAS, the regions not overlapped by HTAP1 and HTAP2 are mostly in the less populated/polluted regions such as the northwestern China, according to the HTAP2 emission maps (http://edgar.jrc.ec.europa.eu/htap_v2/index.php). HTAP2's EUR domain excludes certain regions in Russia/Belarussia/Ukraine, Middle East and 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.

A unitless "Response to Extra-Regional Emission Reductions (RERER)" metric (Galmarini et al., 2017), as defined in eq. (2), was also calculated to measure the importance of local versus non-local sources to NAM's O_3 levels:

$$RERER\ (O_3, NAM) = \frac{R_{O3,non-NAM,20\%}}{R_{O3,global,20\%}} = \frac{(NAMALL\ O3 - GLOALL\ O3)}{(BASE\ O3 - GLOALL\ O3)} \quad (2)$$

The denominator and numerator terms of RERER represent the impacts of global and non-NAM anthropogenic emissions on NAM O_3, respectively. The higher the NAM RERER value is, the stronger impact from non-local sources on NAM is indicated. The RERER value can exceed 1, when emission reductions led to increasing concentrations (e.g. O_3 titration by nitrogen monoxide (NO)).

The STEM (version 2K3) regional simulations were then performed on a 60 km×60 km horizontal resolution (a typical coarse regional model resolution) grid over NAM within the domain defined in Figure 2a during May-June 2010. The meteorological conditions in spring 2010 were compared with the climatology from the NCEP/NCAR reanalysis data for the 1981-2010 period (Kalnay et al., 1996) in Huang et al. (2013b), concluding that this spring represents a period of stronger-than-climatological average spring trans-Pacific transport, based on a stronger meridional gradient in the North Pacific and higher Pacific/North American (PNA) indexes. This is consistent with the findings by Lin et al. (2014) that the El Niño conditions during the 09/10 winter strengthened the trans-Pacific transport of Asian pollution in spring 2010. 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). In contrast, higher-than-normal temperatures were found in the eastern US that favored anomalously strong local O_3 production.

STEM has been used to interpret the observations collected by satellites and during aircraft campaigns in the past decade (e.g., Carmichael et al., 2003a, b; Huang et al., 2010, 2013a, b, 2014, 2015). STEM calculates gas-phase chemistry reactions based on the SAPRC 99 gaseous chemical mechanism (Carter, 2000) with thirty photolysis rates calculated online by the Tropospheric Ultraviolet-Visible radiation model (Madronich et al., 2002). Most of the key configurations of the 60 km base simulations are the same as those described in Lapina et al. (2014), i.e., meteorological

Deleted: -
Deleted: -
Deleted: -
Deleted: -
Deleted: 6
Deleted: 2b
Deleted: S
Deleted: (
Deleted:) and stratospheric intrusion conditions (based on the tropopause pressure and the UTLS relative humidity).
Formatted: Font:Not Italic

fields were pre-calculated by the Advanced Research Weather Research and Forecasting Model (WRF-ARW, Skamarock et al., 2008) version 3.3.1 forced by the North American Regional Reanalysis data (Mesinger et al., 2006), using a similar set of the physics configuration to those in Huang et al. (2013a). Biomass burning emissions are from the Fire INventory from NCAR (FINN) inventory version 1.0 (Wiedinmyer et al., 2011). Biogenic emissions were calculated by the Model of Emissions of Gases and Aerosols from Nature (MEGAN) version 2.1 (Guenther et al., 2012), driven by the WRF meteorology. Lightning NO_x emissions are generated following the method in Allen et al. (2012), with the flash rates determined by the WRF convective precipitation and scaled to the National Lightning Detection Network flash rates. A major difference of the STEM simulations in this study from the Lapina (2014) study is that the anthropogenic emissions were replaced with the monthly-mean HTAP2 inventory with no weekday-weekend variability applied, rather than the earlier National Emission Inventory (NEI) 2005 in which the weekday-weekend variability exists. 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. The VOC speciation for the SPRAC 99 chemical mechanism in the NEI 2005 (ftp://aftp.fsl.noaa.gov/divisions/taq/emissions_data_2005) were applied to break down the total NMVOC emissions provided in the HTAP2 inventory. The VOC speciation based on the year of 2005 can be unrealistic for 2005 as well as 2010 as studies have reported variable temporal changes of different VOC species in some US cities (e.g., Warneke et al., 2012). The time-varying lateral and top boundary conditions in the STEM base simulations were downscaled from three global models (i.e., 3 hourly SNU GEOS-Chem, 3 hourly ECMWF C-IFS, and 6 hourly RAQMS) base simulations. In support of the SR relationship study to quantify the East Asia anthropogenic impacts on the NAM, three STEM sensitivity simulations were also conducted in which the STEM boundary conditions were downscaled from the EASALL(-20%) sensitivity simulations by these three global models (Table 1b). All STEM simulated 3D chemical fields were saved hourly for the convenience of calculating the US primary O_3 standard metric MDA8 as well as the quantitative comparisons against the satellite Level 2 (L2) O_3 products. The STEM base case surface O_3 performance and its O_3 sensitivities were also compared with those of its boundary condition models as well as the multi- global model means. The latitude/longitude ranges (20-50°N/130-65°W) of NAM for the global and regional model based sensitivity calculations were selected to mainly account for the coverage of the STEM domain, which are slightly different from the definition of North America in HTAP1.

Deleted: This change can introduce uncertainty, but

Deleted: .

Formatted: Indent: First line: 0.5"

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.

2.2.2. Additional base and sensitivity simulations from selected models

In addition to the base and 20% EAS all-category emission perturbation simulations, the global RAQMS model conducted a sensitivity simulation in which the East Asian anthropogenic emissions were zeroed out, which was also used as STEM's boundary conditions (Table 1b). We calculate the "S_{O3}" metric (eq. (3)) using the O_3 sensitivities in STEM and RAQMS at the receptor regions in response to both 20% and 100% of emission reductions, to explore the relationships between the O_3 sensitivity and the size of the emission perturbation. A closer-to-one "S_{O3}" value indicates higher scalability of the sensitivity based on the 20% emission perturbation method for obtaining the full "contribution" of the East Asian anthropogenic emissions on the NAM O_3.

$$S_{O3} = R(O_3, EAS, 100\%)/R(O_3, EAS, 20\%)/5 \quad (3)$$

Where: $R(O_3, EAS, 100\%) = BASE\ O_3 - EASALL(-100\%)\ O_3$

The RAQMS model also provided a base simulation that assimilated satellite O_3 products from the Ozone Monitoring Instrument (OMI, Levelt et al., 2006) and Microwave Limb Sounder (MLS, Livesey et al., 2008) (Pierce et al., 2007), which was used to help better understand the regional model base run error sources, as well as for demonstrating the use of satellite observations to help improve the representation of the trans-boundary pollution.

We also used a number of sensitivity simulations produced by the GEOS-Chem adjoint model v35f in which the emissions from selected anthropogenic emission sectors (power&industry, transportation, residential) or individual O_3 precursor chemical species (NO_x, VOC, CO) over the East Asia were reduced by 20%. Additional simulations for the 2008-2009 periods by the SNU GEOS-Chem were also utilized to quantify the East Asia and non-NAM anthropogenic source impacts in comparison with the 2010 conditions that we mainly focus on in this study.

2.3. In-situ and satellite observations

2.3.1. In-situ observations

Over the receptor NAM, the hourly O_3 observations at the Clean Air Status and Trends Network (CASTNET, http://epa.gov/castnet/javaweb/index.html) sites were used to evaluate the global and regional models' base simulations in four subregions: western US (i.e., the EPA regions 8, 9, 10); southern US (i.e., the EPA regions 4 and 6), the Midwest (i.e., the EPA regions 5 and 7), and the northeast (i.e., the EPA regions 1-3). The numbers of sites used in global and regional models' evaluation in each US subregion are summarized in Tables 2-3. The locations of these sites and the subregions they belong to are indicated in Figure 2a, overlaid on a model-based terrain height map. A majority of the CASTNET sites in the western US are located at high elevation (>1 km) remote or rural regions, more susceptible to the trans-boundary pollution (e.g., Jaffe, 2011). Most of the sites in the other three subregions are located in low elevation regions, mainly affected by local and regional pollution. The model-based terrain heights fairly well represent the reality on subregional scale – the differences between the actual and model-based subregional mean terrain heights at the CASTNET sites are smaller than 0.1 km (Table 3).

During May-June 2010, intense ozonesonde measurements were made at multiple California locations (Cooper et al., 2011), in support of the NOAA "California Nexus (CalNex): Research at the Nexus of Air Quality and Climate Change" field experiment (Ryerson et al., 2013). They have been used to evaluate the simulated O_3 vertical profiles by the HTAP2 participating models. The detailed evaluation results have been shown by Cooper et al. (2016), and will be covered by subsequent publications.

Deleted:

Deleted: ... [1]

Deleted: 2b

Formatted: Indent: First line: 0.5"

Over HTAP2's EAS source region, the global models' O_3 performance was evaluated against the monthly-mean surface in-situ O_3 measurements at 11 sites within the Acid Deposition Monitoring Network in East Asia (EANET, http://www.eanet.asia) that had data throughout the year of 2010. These include eight Japanese and three Korean sites (Figure 3a), all of which are located at low elevation regions (2-150 m). The reported monthly mean observations at these sites were based on weekly or daily sampled data, varying among sites.

2.3.2. Satellite products

In two case studies of high O_3 episodes, L2 and L3 O_3 and CO retrievals from several satellite instruments were used to assess the impacts of trans-Pacific pollution transport and stratospheric O_3 intrusions on NAM O_3 levels in early May. These include: 1) the early afternoon O_3 and CO profiles version 5 from the Tropospheric Emission Spectrometer (TES) (Beer et al., 2001; Beer, 2006) on the Aura satellite; 2) the mid-morning O_3 profiles from the METOP-Infrared Atmospheric Sounding Interferometer (IASI), which were retrieved using the Jet Propulsion Laboratory (JPL) TES optimal estimation retrieval algorithm (Bowman et al., 2006) for selected areas including the western US (Oetjen et al., 2014, 2016); as well as 3) the early afternoon L3 O_3 and CO maps (version 6, 1°×1°) from the Aqua Atmospheric Infrared Sounder (AIRS) instrument. The TES tropospheric O_3 retrieval is often sensitive to the mid- to lower free troposphere, and O_3 at these altitudes in the Eastern Pacific is known to possibly impact the downwind US surface air quality at later times (Huang et al., 2010; Parrish et al., 2010). 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). Note that IASI O_3 data are processed operationally in Europe using a different algorithm. For this work we used O_3 profiles from TES and IASI processed using a consistent algorithm at JPL, although the latter set of data represents only a small subset of the full set of the IASI radiance measurements. The IASI and TES L2 O_3 profiles (screened by the retrieval quality and the C-Curve flags) were used to evaluate the STEM O_3 vertical distributions in the different base simulations, and the satellite observation operators were applied in these comparisons. Taking TES as an example, its observation operator h_z for O_3 is written in (4):

$$h_z = z_c + \mathrm{A}_{\mathrm{TES}}\,(\ln(F_{TES}(c)) - z_c) \quad (4)$$

where z_c is the natural log form of the TES constraint vector (a priori) in volume mixing ratio. A_{TES} is the averaging kernel matrix reflecting the sensitivity of retrieval to changes in the true state (Rodgers, 2000). F_{TES} projects the modeled O_3 concentration fields c to the TES grid using spatial and temporal interpolation. The exponential of h_z is then used to compute the mismatches between the model and TES O_3 retrievals as the model evaluation. A small mismatch between model with the satellite observation operators and the satellite retrievals may indicate either good model performance or may be the low sensitivity of the retrievals to the true O_3 profile. AIRS O_3 is sensitive to the altitudes near the tropopause, with positive biases over the ozonesondes in the upper troposphere (e.g., Bian et al., 2007); AIRS CO is most sensitive to 300–600 hPa (Warner et al., 2007) and is frequently used together with the AIRS O_3 to distinguish the stratospheric O_3 intrusions from long-range transported anthropogenic or biomass burning pollution. We use the L3 AIRS products in this study to get a broad overview of the areas that are strongly impacted by the stratospheric O_3 intrusions or/and LRT of pollution.

Formatted: Indent: First line: 0.5"

Deleted: a

Deleted: study

Field Code Changed

The bottom-up NO_x emissions from the HTAP2 inventory were assessed on a monthly base by comparing the GEOS-Chem nitrogen dioxide (NO_2) columns with the de-striped KNMI (Royal Netherlands Meteorological Institute) OMI column NO_2 product version 2.0 (Boersma et al., 2011a, b). For this model evaluation against the OMI L2 products, the NO_2 fields calculated by the GEOS-Chem adjoint model were saved daily at 13:30 local solar time, roughly coinciding with the Aura and Aqua overpassing times. Other parameters used in the model column calculations came from the GEOS-5/GEOS-Chem monthly mean conditions. The OMI data that passed the tropospheric quality flag at 13-14 local time were selected based on the following screening criteria: surface albedo<0.3; cloud fraction<0.2; solar zenith angle <75°; and viewing zenith angle <45°. The averaging kernels (Eskes and Boersma, 2003) and Air Mass Factors (AMFs) in the KNMI product were used to calculate the modeled tropospheric NO_2 vertical columns comparable to the OMI's. Details of the method to compare the model-based NO_2 columns with the KNMI OMI's can be found in Huang et al. (2014).

3. Results and Discussions

3.1. Evaluation of the HTAP2 bottom-up NO_x emissions and the model base simulations

3.1.1. Evaluation of the bottom-up NO_x emissions

The comparison of the GEOS-Chem adjoint NO_2 columns with the OMI product was used to help assess the bottom-up HTAP2 NO_x emissions. Figure 4 shows that NO_2 columns from GEOS-Chem's base simulations over the US are overall overestimated. 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). Larger OMI-model disagreement was found over the central/eastern US in June 2010 than in May, likely also due to the uncertainty in GEOS-Chem's soil or lightning NO_x emissions, which appear to be high over these regions (Figure S1). The NO_2 columns in the GEOS-Chem base simulation were overestimated in many northern China rural areas and underpredicted in a few urban areas in the East Asia as well as a broad area in the southwestern China. The mismatches between model and OMI NO_2 fell within the ranges of the comparison between the GOME2 NO_2 column product and six models' simulations over China in summer 2008 (Quennehen et al., 2016). 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. These global model evaluation results suggest that the EAS-NAM SR relationships analyzed using this inventory may overall overestimate the NAM local contribution and underestimate the EAS contribution—Under different chemical regimes, this statement would also rely on the quality of other O_3 precursors' emissions in the HTAP2 inventory, and they may be associated with variable uncertainties depending on the species or emission sector as introduced in Section 2.1. Therefore, careful assessment of other key O_3 precursors' emissions in the inventory is needed in the future work. 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. We also recommend more global models to save their calculations more frequently, at least near the satellite overpassing

Deleted: p

Deleted: *model base simulations and*

Deleted: global model O_3 ensembles and the

Moved down [2]: The monthly-mean surface O_3 from multiple global models' free runs was evaluated with the CASTNET observations, at the stations with 95% of the hourly O_3 observation completeness for the 1 May-30 June 2010 period

Formatted: No underline

Formatted: Indent: First line: 0.5"

Deleted: , and the mean biases and RMSEs for these two months were summarized in Table 2 by US subregions. The three boundary condition-model as well as the eight-model ensembles overall underpredicted O_3 in the western US (by ~3-6 ppbv), similar to the HTAP1 model performance over these regions for May-June 2001.

Moved down [3]: This can be due to the underestimated trans-boundary pollution (as indicated by the evaluation of modeled O_3 profiles with ozonesondes and satellite O_3 products). In addition, the coarser model resolutions are less capable of resolving the local features that influence the pollutants' import processes, chemical transformation, as well as regional processes such as the cross-state pollution transport over complex terrains. The global RAQMS base simulation with satellite assimilation improved the free tropospheric O_3 structure as its comparisons with the ozonesondes shows, which also enhanced the simulated monthly-mean surface O_3 by up to

Deleted: over 10 ppbv in the western US and some coastal areas in the southeastern US (Figure S1, left). ... [2]

Moved down [4]: subregions (by 8-12 ppbv), close to HTAP1 model performance for May-June 2001 over the similar areas (Fiore et al., 2009) and in the Lapina et al.

Deleted: (2014) study for 2010, in large part due to the ... [3]

Moved down [5]: Satellite assimilation led to 2-6 ppb ... [4]

Moved down [7]: 2016; Travis et al.,

Formatted: Indent: First line: 0"

Deleted: Except in the northeastern US, the eight-mod ... [5]

Moved down [6]: in the literature (e.g., Geddes et al., 2016;

Deleted: 2016), but how serious these issues were in th ... [6]

Formatted: Underline

Deleted: 3

Deleted: , and larger disagreement was found over the ... [7]

Deleted: overall there does appear to be

Deleted: consistent with the findings of

Deleted: It is likely that other O_3 precursors' co-emitte ... [8]

Deleted: This

Deleted: precursors

Deleted: emission

Deleted: so

Deleted: also needed. Note

times, for a more comprehensive assessment of the emission inventory and a better understanding of the model biases.

3.1.2. Evaluation of the global model O_3 performance in NAM and EAS

The monthly-mean surface O_3 from multiple global models' free runs was evaluated with the CASTNET observations, at the stations with 95% of the hourly O_3 observation completeness for the 1 May-30 June 2010 period. The mean biases and RMSEs for these two months were summarized in Table 2a by US subregions. The three boundary condition-model as well as the eight-model ensembles overall underpredicted O_3 in the western US (by ~3-6 ppbv), similar to the HTAP1 model performance over these regions for May-June 2001 presented in Fiore et al. (2009). This can be due to the underestimated trans-boundary pollution (as indicated by the evaluation of modeled O_3 profiles with ozonesondes and satellite O_3 products). In addition, the coarser model resolutions are less capable of resolving the local features that influence the pollutants' import processes, chemical transformation, as well as regional processes such as the cross-state pollution transport over complex terrains. The global RAQMS base simulation with satellite assimilation improved the free tropospheric O_3 structure as its comparisons with the ozonesondes shows, which also enhanced the simulated monthly-mean surface O_3 by up to >10 ppbv in the western US and some coastal areas in the southeastern US (Figure S2, left). The global models overall significantly overestimated O_3 in the other three subregions (by 8-12 ppbv), close to HTAP1 model performance for May-June 2001 over the similar areas (Fiore et al., 2009) and in the Lapina et al. (2014) study for 2010, in large part due to the uncertainties in the bottom-up emissions as discussed in Section 3.1.1. Satellite assimilation led to 2-6 ppbv higher RAQMS surface O_3 in the central/southern/eastern US than in its free simulation, which are associated with higher positive biases.

The surface O_3 performance by individual global models varies significantly, e.g., with the RMSEs at all CASTNET sites ranging from ~9 ppbv to >15 ppbv (Table 2b). 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 individuals. 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.

The monthly-mean surface O_3 from multiple global models' free runs was also evaluated with the EANET observations. Among the three boundary condition models, GEOS-Chem produced higher O_3 than the other two throughout the year, and C-IFS O_3 is the lowest from April to December. The three-model and eight-model ensembles are lower than the surface O_3 observations by <10 ppbv during high O_3 seasons (winter/spring), but show substantial (>10 ppbv) positive biases during low O_3 seasons especially in July and August (Figure 3b), similar to the HTAP1 model performance over Japan in 2001 (Fiore et al., 2009). During May-June 2010,

Deleted:
Moved (insertion) [2]
Formatted: No underline
Formatted: Indent: First line: 0.5"
Moved (insertion) [3]
Moved (insertion) [4]
Moved (insertion) [5]
Moved (insertion) [6]
Moved (insertion) [7]

generally the models performed better at the Japanese sites than at the Korean sites (Table 2c), with significant positive biases occurring at low O_3 regions (e.g., in central Japan) and negative biases found at high O_3 regions, mainly owing to the uncertainty in the local and upwind emissions. The different approaches to generate the monthly-mean modeled and the observed O_3 data may have also contributed to these model-observation discrepancies. Overall O_3 performance by individual models varies less significantly than at the CASTNET sites, with RMSEs ranging from 8.6 ppbv to ~13 ppbv (Table 2b). The three-model ensemble outperforms two individual models, and the eight-model ensemble outperforms six individual models. Unlike at the CASTNET sites, the three-model ensemble agrees better with the observations than the eight-model ensemble (Table 2c).

3.1.3. Evaluation of the STEM regional base simulations w/ three sets of boundary conditions

The three STEM base simulations using different boundary conditions were evaluated with the hourly O_3 observations at the CASTNET sites in the four US subregions. The evaluation included the 8 May-30 June 2010 period to exclude the results during the one-week spin-up period. The time series plots of observed and modeled O_3 at the western US CASTNET sites show that STEM was capable of capturing several high O_3 periods, and it produced larger biases during the nighttime (Figure 2c), as a result of the poorer WRF performance. Figure 2c and the evaluation statistics in Table 3a-b indicate that STEM/C-IFS O_3 concentrations are associated with the highest positive bias and RMSE, while the STEM/GEOS-Chem and STEM/RAQMS predictions were positively and negatively biased by less than 2 ppbv, respectively, with similar RMSEs and correlations with the observations. The quality of the three STEM simulation mean is closest to the STEM/GEOS-Chem run, with the mean bias/RMSE of ~1.6/4.9 ppbv, much better than the three-boundary model ensemble (-5.7/10.4 ppbv). However, this good performance can be a net effect of incorrect partitioning between the trans-boundary and local source contributions, with the former being underestimated and offsetting the overestimation of the latter. 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).

The three STEM base simulations all significantly overpredicted O_3 over the rest of the US in part due to the overall overestimated NO_x emissions, with the STEM/RAQMS associated with the lowest RMSEs and mean biases, but STEM/C-IFS correlated best with the observations (Table 3b). These positive biases are higher than the global model ensembles', which can partially result from the possible unrealistic VOC speciation of the emission inventory and the SAPRC 99 chemical mechanism: Although SAPRC mechanisms have been used in air quality modeling for regulatory applications in some US states such as California, they usually produced higher O_3 than other mechanisms such as the CB04 and the CB05 (which were used by some HTAP2 global models, see Table 1c) over the US, and the comparisons between SAPRC 99 and SAPRC 2007 are still in progress (e.g., Luecken et al., 2008; Zhang et al., 2012; Cai et al., 2011). It is important to timely update the chemical mechanisms in the chemistry models, and we also suggest to timely upgrade the VOC speciation in the bottom-up emission inventories in the US to benefit the air

Deleted: averaged and

Deleted: The evaluation statistics is summarized in Table 3. The time series of observed and modeled O_3 at the western US CASTNET sites are shown in Figure 2a where the model overall simulated the surface O_3 fairly well, with a much smaller mean bias (~1.6 ppbv) than the global model ensembles.

Deleted: S1

Deleted: model

Deleted: mainly

Deleted: .

Deleted: CB05

quality modeling. Additionally, the uncertainty from non-anthropogenic emissions, such as the biogenic VOC emissions from WRF/MEGAN which is known to often have positive biases, can be another cause: As Hogrefe et al. (2011) presented, the MEGAN emissions resulted in a higher O_3 response to hypothetical anthropogenic NO_x emission reductions compared with another set of biogenic emission input. 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. These temperature and radiation biases, can also be important sources of uncertainty in the modeled O_3 production. 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. 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). We anticipate that the results from the Air Quality Model Evaluation International Initiative (AQMEII) experiment (e.g., Schere et al., 2012; Solazzo et al., 2012; Galmarini et al., 2015, 2017), which involves more regional model simulations over the US with the similar set of boundary conditions but different chemical mechanisms and non-anthropogenic emission inputs, can help better understand the causes of errors in the simulated total O_3.

3.2. The NAM surface O_3 sensitivity to extra-regional anthropogenic pollutants

3.2.1. Global model ensembles

The impact of all foreign (i.e. non-NAM) anthropogenic sources on NAM surface O_3 was first explored, including the spatial distributions of the RERER metric (eq. (2)) based on various global models' simulations (Figure 5), and the domain wide mean sensitivities R (O_3, non-NAM, 20%) (eq. (1d)) (Figure 6). Across the NAM, the strongest impacts were found in spring time (March-April-May, larger than 1.5 ppbv in average over the domain) and the weakest impacts are shown during the summertime (June-July-August, 1.0-1.3 ppbv), consistent with the existing knowledge on the seasonal variability of the non-local pollution impacts on NAM for other years (e.g., Fiore et al., 2009; Reidmiller et al., 2009). All global models indicate strong non-NAM anthropogenic source impacts on the western US mainly due to the impact of its high elevation, and also near the US-Mexico border areas, especially southern Texas, due to their vicinity to the Mexican emission sources. Over the western states, stronger non-local impacts were reflected from the results based on higher-horizontal resolution global models (e.g., the >0.6 RERER values from the half degree EMEP model, corresponding to its higher R(O_3, non-NAM, 20%) values than the other models'), similar to the findings in previous modeling studies (Lin et al., 2010, 2012a). Although on a coarse horizontal resolution of 2.8°, OsloCTM3 suggests stronger extra-regional source influences on the northwestern US and the US-Canada border regions than the other models. Its largest number of vertical layers among all global models might be a cause. Larger-than-1 RERER values are often seen near the urban areas and large point sources due to the titration, especially evident from the higher resolution model results. The R(O_3, EAS, 20%) values are larger than 1/3 of the R(O_3, non-NAM, 20%) (0.2-0.5 ppbv from April to June), more than 3-4 times higher than R(O_3, EUR, 20%) and R(O_3, SAS, 20%). Note that all eight models contributed to the R(O_3, EAS, 20%) calculations, but one or two models did not provide all necessary sensitivity runs to compute the RERER, R(O_3, non-NAM, 20%), R(O_3, EUR, 20%), or R(O_3, SAS, 20%).

Deleted: Some factors that caused the overpredicted MEGAN emissions, such as positively-biased temperature in WRF, can also be important sources of uncertainty in the STEM modeled O_3.

Deleted: 6

Deleted: other

Formatted: Indent: First line: 0.5"

Deleted: 4

Deleted: 5

Deleted: ; Brown-Steiner and Hess, 2011).

Comparing to the HTAP1 modeling results, the magnitudes of R(O_3, EUR, 20%) from this study are smaller by a factor of 2-3; In contrast, the R(O_3, non-NAM, 20%) and R(O_3, EAS, 20%) values are >50% higher than the HTAP1 modeling results. The different HTAP1 and HTAP2 results are possibly due to the following three reasons: 1) 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; 2) the changes in the HTAP2 experiment setup from HTAP1. This includes the differences in the participating models, and the different region definitions, 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. For EAS and SAS, however, the regions not overlapped by HTAP1 and HTAP2 are mostly in the less populated/polluted regions; 3) the stronger-than-normal transport in 2010 than in 2000-2001, as first introduced in Section 2.2.1. Interannual variability of R(O_3, EAS, 20%) and R(O_3, non-NAM, 20%) is also found between 2010 and 2008-2009, based on the SNU GEOS-Chem calculations (Figure S3). Foreign anthropogenic pollution impact on NAM was stronger in 2010 than in 2008-2009, especially in April-May. This can be in part due to the higher O_3 precursors' emissions in 2010 from extra-regions including the East Asia (Table S1), as well as the spring 2010 meteorological conditions that favored the trans-Pacific pollution transport.

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. 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. Figure 6b compares the R(O_3, EAS, 20%) estimated by individual boundary condition models, their ensemble mean sensitivities, and the eight-global model mean. The averaged R(O_3, EAS, 20%) from the boundary condition model results are smaller than the eight-global model mean, and except for July-October 2010, GEOS-Chem gives higher R(O_3, EAS, 20%) than RAQMS and C-IFS, consistent with its highest O_3 prediction in the EAS source region (Figure 3b). 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.

The O_3 sensitivities in response to the perturbations of individual species or sector emissions in East Asia, estimated by the GEOS-Chem adjoint model, were also analyzed (Figure S3). These sensitivities show similar seasonal variability to R(O_3, EAS, 20%), with the values ~twice as high in the spring than in summer, also consistent with the results on previous years based on the 20% emission perturbation approach (e.g., Fiore et al., 2009; Brown-Steiner and Hess, 2011; Emmons et al., 2012). However, this seasonal variability is weaker than the results based on the tagged tracer approach for earlier years: Using the CAM-Chem model, Brown-Steiner and Hess (2011) reported that during the springtime, Asian O_3 created from the anthropogenic/biofuel NO_x emissions affected NAM O_3 ~three times as strongly as in summer. This is because the nonlinear O_3 chemistry, which is stronger outside of summer, caused larger O_3 responses to a 100% reduction of NO_x emissions than 5 times of the O_3 responses to a 20% reduction of NO_x emissions.

Deleted: are smaller by a factor of 2-3, as a result of the substantial improvement in the European air quality over the past decades (Crippa et al., 2016; Pouliot et al., 2015), ... [9]
Deleted: The R(O_3, EAS, 20%) based on the emissio ... [10]
Deleted: . The SNU GEOS-Chem-based R(O_3, EAS, ... [11]
Deleted:). Such interannual variability can also be due to
Deleted: S
Deleted: , as introduced in Section 2.2.1
Formatted: Font color: Text 1
Deleted: 5b
Formatted: Font color: Text 1
Deleted:
Formatted: Font color: Text 1
Formatted: Font color: Text 1
Deleted:
Formatted: Font color: Text 1
Deleted: except for July-October 2010. The
Formatted: Font color: Text 1
Deleted:
Formatted: Font color: Text 1
Deleted: the
Formatted: Font color: Text 1
Deleted: overall
Formatted: Font color: Text 1
Formatted: Font color: Text 1
Deleted: (<<5%) and their biases
Formatted: Font color: Text 1
Deleted: , as the impact of the EAS anthropogenic so ... [12]
Formatted: Font color: Text 1
Formatted: Font color: Text 1
Deleted: s
Formatted: Font color: Text 1
Deleted: controlling
Formatted: Font color: Text 1
Deleted: local and regional emissions
Formatted: Font color: Text 1
Deleted: still
Formatted: Font color: Text 1
Deleted: more effective
Formatted: Font color: Text 1
Deleted: complying with
Formatted: Font color: Text 1
Deleted: tighter air quality standard
Formatted: Font color: Text 1
Deleted: S2).

The EAS anthropogenic NO_x emissions more strongly impacted the NAM surface O_3 than the other major O_3 precursors, similar to the findings in Fiore et al. (2009) and Reidmiller et al. (2009) using the perturbation approach, as well as the conclusions in Lapina et al. (2014) based on the adjoint sensitivity analyses. Emissions from the power&industrial sectors are higher in East Asia than the other sectors (Table S1), resulting in its stronger influences on the NAM surface O_3. As the observed NO_2 columns started to drop since 2010 due to the effective denitration devices implemented at the Chinese power and industrial plants (e.g., Liu et al., 2016), depending on the changes in the VOC emissions, it is anticipated to see different R(O_3, EAS, 20%) values for the years after 2010. Therefore, continued studies to assess the East Asian anthropogenic pollution impacts on NAM during more recent years is needed. As emissions from various source sectors can differ by their emitted altitudes and temporal (from diurnal to seasonal) profiles, efforts should also be placed to have the models timely update the heights and temporal profiles of the emissions from those various sectors.

Deleted:

Deleted: .

Deleted: E

3.2.2. Regional model sensitivities and their connections with the boundary condition models'

The monthly-mean STEM surface R(O_3, EAS, 20%) sensitivities based on different boundary condition models were inter-compared, and also compared with the R(O_3, EAS, 20%) estimated by their boundary condition models as well as the global model ensemble mean (Figure 7). For both May and June 2010, the domain-wide mean R(O_3, EAS, 20%) values from STEM/RAQMS were higher than the estimates from RAQMS by 0.03 ppbv; the STEM/GEOS-Chem R(O_3, EAS, 20%) values are lower than those of GEOS-Chem by 0.01-0.06 ppbv, and the STEM/C-IFS R(O_3, EAS, 20%) is 0.02 ppbv higher than C-IFS's in June but slightly (<<0.01 ppbv) lower in May. These differences are overall smaller than the inter-global model differences, and can be due to various factors including the uncertainties in boundary condition chemical species mapping, and the different meteorological/terrain fields/chemistry in the global and regional model pairs. The STEM R(O_3, EAS, 20%) ensemble mean values, however, are less than 0.02 ppbv different from its boundary condition model's ensemble mean for both months. The STEM R(O_3, EAS, 20%) ensemble mean value in June is also close to the eight-global model ensemble mean, but is ~0.05 ppbv lower than the eight-model mean in May. Choosing other/more global model outputs as STEM's boundary conditions may lead to different STEM ensemble mean R(O_3, EAS, 20%) estimates. We also found that the period mean R(O_3, EAS, 20%) of ~0.2 ppbv sampled only at the CASTNET sites (Table 3a) are smaller than those averaged in all model grids. This indicates that currently the sparsely distributed surface network (especially over the western US that is more strongly affected by the extra-regional sources than the other US regions) may miss many LRT episodes that impact the NAM. The planned geostationary satellites with ~2-5 km footprint sizes and hourly sampling frequency (Hilsenrath and Chance, 2013) will help better capture the high O_3 and LRT episodes in these regions.

Deleted: 6

Deleted: CIFS

Deleted: models

Deleted: 3

The spatial patterns of the monthly-mean STEM surface R(O_3, EAS, 20%) sensitivities based on the three boundary condition models are notably different, but overall resemble what's estimated by the corresponding boundary condition model, and the STEM sensitivities show more local details in certain high elevation regions in the US west (Figure 8 shows the June 2010 conditions as an example). These different sensitivities were investigated further, by examining the R(O_3, EAS, 20%) values near the source regions (i.e., East Asia) as well as near the receptor regions (Figure 9). More East Asian anthropogenic O_3 seems to be transported at the upper

Deleted: 7

Deleted: 8

troposphere in RAQMS than in the other two models. GEOS-Chem and RAQMS R(O_3, EAS, 20%) sensitivities are similar over the EAS as well as the 500-900 hPa near the receptor in the eastern Pacific (at ~135°W), the altitudes US surface O_3 are most strongly sensitive to during the summertime as concluded from previous studies (e.g., Huang et al., 2010, 2013a; Parrish et al., 2010). Despite the close NAM domain-wide mean values from the STEM/GEOS-Chem and STEM/RAQMS, the spatial patterns of R(O_3, EAS, 20%) over NAM differ in these two cases, with the latter case showing sharper gradients especially in the western US, partially due to the impact of its higher horizontal resolution. The R(O_3, EAS, 20%) values from STEM/C-IFS are lower than from the other two cases both near the sources and at (near) NAM. The STEM surface (also near surface, not shown in figures) R(O_3, EAS, 20%) does not spatially correlate well with the column R(O_3, EAS, 20%), the latter of which contributed more to the base case O_3 columns, indicating that a good portion of the transported East Asian pollution did not descend to the lower altitudes to impact the boundary layer/ground level air quality. An additional regional simulation was performed in which the STEM boundary conditions were downscaled from a RAQMS simulation without the East Asian anthropogenic emissions. The non-linear emission perturbation-O_3 response relationships, as the larger-than-1 S_{o3} metric (eq. (3)) indicate, are seen across the domain, for both the surface and column O_3 (Figure 8). S_{o3} for column O_3, ranging from 1.15-1.25 in most regions, are overall ~0.05 higher than S_{o3} for the surface O_3. Therefore, the full source contribution obtained by linearly scaling the receptor regional mean O_3 sensitivity to the 20% reduction in the source region emissions may be underestimated by at least ~10%.

3.2.3. Regional model MDA8 sensitivities on all days and during the O_3 exceedances

The temporal variability of the STEM R(O_3, EAS, 20%) ensemble sensitivities were also studied. For most US subregions, 3-6 LRT episodes (defined as when the sensitivities are above the period mean) were identified during May-June. Throughout this period, the hourly R(O_3, EAS, 20%) and the observed O_3 at the surface CASTNET sites are weakly correlated (Table 3a), but they display similar diurnal cycles (e.g., Figures 2c and 2d for the western US sites), possibly because the deeper boundary layer depth during the daytime enhanced entrainment down-mixing of the extra-regional pollutants to the surface. The identified diurnal variability of the R(O_3, EAS, 20%) can cause differences in the calculated MDA8 and all-hour mean R(O_3, EAS, 20%) values. Figure S4 shows that the mean R(MDA8, EAS, 20%) values, usually at daytimes, are higher than the all-hour averaged R(O_3, EAS, 20%) in most STEM model grids during both months. 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 O_3 sensitivities. Also, the hourly sampling frequency of the planned geostationary satellites is anticipated to be more helpful for evaluating the impacts of the LRT episodes.

The STEM R(MDA8, EAS, 20%) in all model grids within the four US subregions were averaged on all days during May-June 2010 and only on the days when the simulated total MDA8 O_3 is over 70 ppbv (Figure 10). These sensitivities also show appreciable spatial variability: from 0.35-0.58 ppbv in the western US (also with the largest standard deviations, not shown), which is slightly higher than the HTAP1 results reported by Reidmiller et al. (2009) for Spring 2001, to ~0.1-0.25 ppbv in the rest three subregions, which is close to the Reidmiller et al. (2009) results.

Comparing the solid bar plots in Figures 10-11, we found that on all days in the three non-western subregions, R(MDA8, EAS, 20%) values sampled at CASTNET sites are slightly smaller

Deleted: CIFS

Deleted: 7).

Deleted: .

Formatted: Indent: First line: 0"

Formatted: Underline

Deleted: 3

Deleted: Figure 2a

Deleted: S3

Deleted: for May-June 2010 in

Deleted: 9

Deleted: Qualitatively consistent with the findings in Reidmiller et al. (2009), R(MDA8, EAS, 20%) is smaller during the high O_3 total days in all subregions. Note that the STEM base simulations overall substantially overpredicted the total O_3 in non-western US regions, so the R(MDA8, EAS, 20%) calculated during the days of O_3 exceedances can actually represent the sensitivities during the non-exceedances.

Formatted: Normal

than those computed for all model grids, while in the non-western states the opposite differences are seen. This again suggests that expanding observation network would help better capture the high O_3 and LRT episodes.

Figure 10 suggests smaller R(MDA8, EAS, 20%) values during the high O_3 days in all subregions. However, STEM's total O_3 concentrations at CASTNET sites during the O_3 exceedances were substantially overpredicted in non-western US regions while significantly underpredicted in the western US (see mean biases above the bar plots in Figure 11). Therefore, the R(MDA8, EAS, 20%) values shown in Figure 10 during O_3 exceedances can actually represent the sensitivities during the non-exceedances in non-western US regions, and may not represent the sensitivities during all O_3 exceedances in the western US. Figures 11-12 show that if calculated only at the CASTNET sites during the exceedances, in non-western US regions, R(MDA8, EAS, 20%) is 0.02-0.07 ppbv smaller during the high O_3 total days. This is qualitatively consistent with the findings in Reidmiller et al. (2009), and is possibly because that the LRT impacts were stronger on some days with good dispersion conditions when the NAAQS was not exceeded, but weaker on some high O_3 days under stagnant conditions. In contrast, western US R(MDA8, EAS, 20%) at CASTNET sites was ~0.05 ppbv higher on high O_3 days than for all days, and this differences are larger in rural/remote areas where local influences are less dominant. As a result, the medium/strong positive correlations are found between modeled LRT of pollution and the total O_3 in these regions (Table 3a; Lin et al., 2012a).

3.3. Case studies of spring (9 May) and summer (10 June) LRT events mixed with stratospheric O_3 intrusions

Lin et al. (2012a, b) and Neuman et al. (2012) showed that the trans-Pacific pollution transport intensely impacted the western US during 8-10 May, 2010, intermingled with a stratospheric intrusion that contributed to at least 1/3 of the total O_3 in some high elevation regions. This episode is indeed indicated by the O_3 and CO products from AIRS and TES at ~500 hPa over the Eastern Pacific (Figure 13), and the observed TES and IASI O_3 profiles over the western US indicated elevated O_3 levels (>80 ppbv) at 700-900 hPa. Huang et al. (2013b) found that the meteorological conditions during this period (i.e., a strong jet at ~700 hPa with wind speed >20 m/s shifted southwesterly when passing the southern California and continued to travel towards the mountain states), along with the orographic lifting, efficiently exported the southern California anthropogenic pollution, which was chemically coupled with the extra-regional pollution and significantly enhanced the O_3 levels in the US intermountain west.

We selected this episode to compare the STEM surface total O_3 concentrations as well as the R(O_3, EAS, 20%) sensitivities based on the different HTAP2 boundary condition models. Figure 14 evaluates the simulated O_3 profiles in the western US from several STEM base simulations against the TES and IASI O_3 retrievals, and Figures 15a-d indicate the performance of the daily surface total MDA8 O_3 from these simulations. We found that the underestimated free tropospheric O_3 from the STEM simulations that used any single free-running chemical boundary conditions contributed to the underestimated STEM surface O_3 in the high elevation mountain states: e.g., by 9-14 ppbv at three CASTNET sites (Grand Canyon National Park (NP), AZ; Canyonlands NP, UT; and Rocky Mountain NP, CO) where O_3 exceedances were observed. The unsatisfactory performance by free-running global models during high O_3 events would pose

Deleted: *y*
Deleted: *the*
Deleted: *2010*
Formatted: Font:Not Italic, Underline
Formatted: Normal, Indent: Left: 0"
Deleted: .
Deleted: E
Deleted: 10
Deleted: 11
Deleted: 2

difficulties for regional models (regardless of their resolutions and other configurations, parameterization) to accurately estimate the SR relationships using boundary conditions downscaled from these model runs. The STEM base simulation using the RAQMS assimilated fields as the boundary conditions, agrees most with the observed O_3 at the CASTNET sites, as well as the TES and IASI O_3 profiles in the western states. Similar to the conclusions drawn in Huang et al. (2010, 2015) for summer 2008, we again demonstrated the robustness of satellite chemical data assimilation for improving the boundary condition models' O_3 performance. 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.

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 R(O_3, EAS, 20%) sensitivities, which also show notable intermodel differences: The estimated R(MDA8, EAS, 20%) in the different STEM cases range from <1.0 ppbv to ~1.3 ppbv, at least 40% higher than the May-June period mean in Figures 10-11. The mean R(MDA8, EAS, 20%) at three high O_3 CASTNET sites range from 0.73 (STEM/GEOS-Chem) to 0.98 ppbv (STEM/C-IFS), with the mean S_{O3} of ~1.14 at these sites based on the STEM/RAQMS runs due to the nonlinear emission perturbation-O_3 response relationships (Figure 15e-h). The R(MDA8, EAS, 100%) from the STEM/RAQMS case is as high as >7 ppbv over the high terrain regions. These are of smaller magnitudes than the estimates in Lin et al. (2012a), possibly due to the differences in the used models and the bottom-up emission inputs.

A stratospheric O_3 intrusion also affected the NE US on the same day, as revealed by the satellite free tropospheric O_3 and CO observations (Figure 13). This intrusion was mixed with LRT East Asian pollution (Figure 15 and Figure S5). However, this intrusion did not enhance the NE boundary layer/surface O_3 concentrations, which were actually anomalously low (MDA8<40 ppbv) as indicated by the model base simulations and the CASTNET observations (Figure 15a-d). Similar characteristics during summertime stratospheric O_3 intrusion events around this region have been discussed by Ott et al. (2016). The East Asian pollution less intensely (<50%) affected the surface O_3 levels in these regions than in the US west, due to the greater transport distances, stronger local emission influence on chemical production/loss, as well as the impact of the overall flat terrain in the US east.

A summertime LRT event on 9-10 June is analyzed to contrast with the 9 May case study. Lin et al. (2012b) showed that >80 ppbv of ozonesonde data in northern California at 2-6 km measured the stratospheric O_3 remnants during this episode, and the transported stratospheric O_3 contributed to as much as ~50% of the total O_3 in southern California based on their model calculations. We show that on 10 June over 100 ppbv of O_3, as well as <90 ppbv CO, was observed by satellites at ~500 hPa above Nevada and northern California (Figure 16), which again was substantially underestimated by all free-running models (Figure 17), resulting in the underpredicted total O_3 at two CASTNET sites in southern California (Converse Station and Joshua Tree NP) that experienced O_3 exceedances on this day (Figure 18a-c). The negative biases in the "transported background" O_3 and surface MDA8 O_3 were successfully reduced by incorporating satellite data (Figures 17 and 18d).

Deleted: .

Deleted: Figure 9. Strong

Deleted: are also shown during this period

Deleted: 12, lower

Deleted: 10

Deleted: 12

Deleted: S4

Deleted: 2

Figures 18e-h show that LRT of EAS anthropogenic pollution also strongly affected southern California and Nevada. Notable intermodel differences are again found in the estimated R(MDA8, EAS, 20%), but they are overall lower than on 9 May (<1.0 ppbv). The mean R(MDA8, EAS, 20%) at the two high O_3 CASTNET sites range from 0.54 (STEM/C-IFS) to 0.86 ppbv (STEM/RAQMS), with the mean S_{O3} of ~1.13 at these sites based on the STEM/RAQMS runs (Figure 18e-h). The R(MDA8, EAS, 100%) from the STEM/RAQMS case is as high as >6 ppbv over southern California and Nevada. Compared to the spring event, R(MDA8, EAS, 20%) in the eastern US are discernable only over a limited region, due to weaker transport and stronger local chemical production/loss.

4. Conclusions and suggestions on future directions

In support of the HTAP Phase 2 experiment that involved high-resolution global models and regional models' participation to advance the understanding of the pollutants' SR relationships in the Northern Hemisphere, we conducted a number of regional scale STEM base and forward sensitivity simulations over NAM during May-June 2010. STEM's top and lateral chemical boundary conditions were downscaled from three global models' (i.e., GEOS-Chem, RAQMS, and ECMWF C-IFS) base and sensitivity simulations (in which the East Asian anthropogenic emissions were reduced by 20%). Despite dilution along the great transport distance, the East Asian anthropogenic sources still had distinguishable impact on the NAM surface O_3, with the period-mean NAM O_3 sensitivities to a 20% reduction of the East Asian anthropogenic emissions (i.e., R(O_3, EAS, 20%)) ranging from ~0.24 ppbv (STEM/C-IFS) to ~0.34 ppbv (STEM/RAQMS). The spatial patterns of the STEM surface O_3 sensitivities over NAM overall resembled those from its corresponding boundary condition model, with regional/period mean R(O_3, EAS, 20%) differed slightly (<10%) from its corresponding boundary condition model's, which are smaller than those among its boundary condition models. The boundary condition models' two-month mean R(O_3, EAS, 20%) was ~8% lower than the mean sensitivity estimated by multiple global models. Therefore, choosing other global model outputs as STEM's boundary conditions may lead to different STEM O_3 sensitivities. The biases and RMSEs in the simulated total O_3, which differed significantly among models, can partially be due to the uncertainty in the bottom-up NO_x emission inputs according to the model comparison with the OMI NO_2 columns, and future work on attributing the intermodel differences on process level is particularly important for better understanding the sources of uncertainties in the modeled total O_3 and its source contribution.

The HTAP2 multi-model ensemble mean R(O_3, EAS, 20%) values in 2010 were higher than the HTAP1 reported 2001 conditions, due to the impacts of the growing East Asian anthropogenic emissions, the interannual variability in atmospheric circulation (i.e., stronger trans-Pacific transport in spring 2010 following an El Niño event), and the different experiment designs of HTAP1 and HTAP2. The GEOS-Chem O_3 sensitivities in 2010 were also higher than the 2008-2009 conditions due to the increasing Asian emissions and the spring 2010 meteorological conditions that favored the trans-Pacific pollution transport. The GEOS-Chem sensitivity calculations indicate that the East Asian anthropogenic NO_x emissions mattered more than the other East Asian O_3 precursors to the NAM O_3, qualitatively consistent with previous adjoint sensitivity calculations. Continued research is needed on temporal changes of emissions for different species and sectors in NAM and foreign countries as well as their impacts on the SR relationships. As emissions from various source sectors can differ by emitted altitudes and

Formatted: Indent: First line: 0"
Deleted: North America
Deleted: The STEM
Deleted: CIFS
Deleted: The STEM
Formatted: Font color: Text 1
Deleted: (including the 24h mean and the policy-relevant MDA8 metric averaged throughout the study period and during a selected transport event) over North America
Deleted: the
Deleted: but can be quantitatively different from the mean sensitivities
Deleted: all
Deleted: model ensembles
Deleted: /more
Deleted: ensemble mean
Formatted: Not Superscript/ Subscript
Deleted: Overall, the monthly-based US O_3 sensitivities to the 20% reduction of the East Asian anthropogenic emissions contributed to <<5% of the total O_3 and are of smaller magnitudes than the
Deleted: in the modeled total O_3. Better quantifying the contributions from other factors, such as the stratospheric O_3 intrusion and the local O_3 formation, would still be the most effective way to help reduce the North American pollution levels and the model uncertainties. The US O_3 sensitivities to the East Asian anthropogenic emissions were episodically strong, contributing to the O_3 exceedances in some high terrain areas. Assessing the sources of
Deleted: are
Deleted: evaluating the East Asian pollution impacts during these episodes. The STEM O_3 sensitivities followed similar diurnal cycles as the total O_3, emphasizing the importance of saving model results hourly for continently calculate policy-relevant metrics, as well as the usefulness of hourly sampling frequency of the planned geostationary satellites for better evaluating the impacts of the LRT events.
Deleted: sensitivities
Deleted: as well as 1/5 of the original estimates for 2001-2005 using the tagged tracers. This indicates the increasing
Deleted: pollution on North America.
Deleted: S
Deleted: North American
Deleted: North America

temporal profiles, efforts should also be placed to have the models timely update the height and temporal profiles of the emissions from various sectors.

An additional STEM simulation was performed in which the boundary conditions were downscaled from a RAQMS simulation without East Asian anthropogenic emissions (i.e., a 100% emission reduction), to assess the scalability of the mean O_3 sensitivities to the size of the emission perturbation. The scalability was found to be spatially varying, ranging from 1.15-1.25 for column O_3 in most US regions, which were overall ~0.05 higher than the surface O_3's. Therefore, the full source contribution obtained by linearly scaling the NAM regional mean O_3 sensitivity to the 20% reduction in the East Asian emissions may be underestimated by at least 10%. The underestimation in other seasons of the HTAP2 study period may be higher and will need to be quantified in future work. Also, 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.

The STEM O_3 sensitivities to the East Asian anthropogenic emissions (based on three boundary condition models separately and averagely) were strong during 3-6 episodes in May-June 2010, following similar diurnal cycles as the total O_3. Stronger-than-normal East Asian anthropogenic pollution impacts were estimated during O_3 exceedances in the western US, especially over the high terrain rural/remote areas; in contrast, non-local pollution impacts were less strong during O_3 exceedances in other US regions. We emphasized the importance of saving model results hourly for continently calculate policy-relevant metrics, as well as the usefulness of hourly sampling frequency of the planned geostationary satellites for better evaluating the impacts of the LRT events.

Based on model calculations, satellite O_3 (TES, JPL-IASI, and AIRS), CO (TES and AIRS) and surface O_3 observations on 9 May 2010, we showed the different influences from stratospheric O_3 intrusions along with the transported East Asian pollution on O_3 in the western and the eastern US. This event was further compared with a summer event of 10 June 2010. During both events, the unsatisfactory performance of free-running global models would pose difficulties for regional models (regardless of their resolutions and other configurations, parameterization) to accurately simulate the surface O_3 and its source contribution using boundary conditions downscaled from these model runs. Incorporating satellite (OMI and MLS) O_3 data effectively improved the modeled O_3. 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. Meanwhile, efforts should also be devoted to advancing and applying higher-resolution regional scale modeling and chemical data assimilation. Furthermore, although satellite observations have been applied for improving the estimated US background O_3 (e.g., Huang et al., 2015), using satellite (and/or other types of)

Deleted: .

Deleted: E

Deleted: and

Deleted: . Motivated

Deleted: Satellite NO_2 (KNMI OMI) and O_3 (TES, JPL-IASI, OMI, MLS, and AIRS) products helped detect pollution episodes, quantify or/and reduce the uncertainties in the bottom-up NO_x emissions and the model transported background O_3. Based on model calculations and satellite/surface observations on a selected day of

Deleted: Continued studies on exceptional events during other seasons are in progress.

Deleted: 6

observations to improve SR relationship studies also needs to be explored. Some of the possible methods include: 1) The combination of data assimilation and the tagging approach; 2) Introducing observation-constrained emission estimates in the emission perturbation analyses.

Acknowledgements

The global and regional modeling results used in this study have been submitted to the AeroCom database following the HTAP2 data submission guidelines (http://iek8wikis.iek.fz-juelich.de/HTAPWiki/HTAP-2-data-submission), or can be made available upon request. Technical support from Anna Carlin Benedictow, Brigitte Koffi, Jan Griesfeller, and Michael Schulz regarding formatting and submitting the data to the AeroCom is acknowledged. MH thanks the research resources at the University of Iowa and JPL/Caltech that supported this study, as well as the travel funding from the US EPA for attending the related HTAP2 workshops. DKH and YD recognize support from NASA AQAST. FD Acknowledges support from the Administrative Arrangement. Part of this research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under contract to the National Aeronautics and Space Administration. Reference herein to any specific commercial product, process or service by trade name, trademark, manufacturer or otherwise does not constitute or imply its endorsement by the United States Government or the Jet Propulsion Laboratory, California Institute of Technology. The views, opinions, and findings contained in this report are those of the author(s) and should not be construed as an official National Oceanic and Atmospheric Administration or U.S. Government position, policy, or decision. We also acknowledge the feedbacks from Dr. Gail Tonnesen, two anonymous reviewers, and Dr. Meiyun Lin on earlier versions of this paper, that helped improve its quality.

Formatted: None, Indent: First line: 0.5", Don't suppress line numbers

Deleted: ... [13]

References

Anderson, D. C., Loughner, C. P., Diskin, G., Weinheimer, A., Canty, T., P., Salawitch, R. J., Worden, H. M., Fried, A., 25 Mikoviny, T., Wisthaler, A., and Dickerson, R., R. (2014), Measured and modeled CO and NO_y in DISCOVER-AQ: An evaluation of emissions and chemistry over the eastern US, Atmos. Environ., 96, 78-87, doi: 10.1016/j.atmosenv.2014.07.004.

Allen, D. J., Pickering, K. E., Pinder, R. W., Henderson, B. H., Appel, K. W., and Prados, A. (2012), Impact of lightning-NO on eastern United States photochemistry during the summer of 2006 as determined using the CMAQ model, Atmos. Chem. Phys., 12, 1737-1758, doi: 10.5194/acp-12-1737-2012.

Ambrose, J.L., Reidmiller, D.R., and Jaffe, D.A. (2011), Causes of high O_3 in the lower free troposphere over the Pacific Northwest as observed at the Mt. Bachelor Observatory. Atmos. Environ., 45, 5302–5315, doi: 10.1016/j.atmosenv.2011.06.056.

Anenberg, S. C., L. W. Horowitz, D. Q. Tong, and J. J. West (2010), An estimate of the global burden of anthropogenic ozone and fine particulate matter on premature human mortality using atmospheric modeling, Environ. Health Perspect., 118(9), 1189–1195.

Avnery, S, D.L. Mauzerall, J. Liu, L.W. Horowitz (2011a), Global Crop Yield Reductions due to Surface Ozone Exposure: 1. Year 2000 Crop Production Losses and Economic Damage, Atmos. Environ., 45, 2284-2296.

Avnery, S, D.L. Mauzerall, J. Liu, L.W. Horowitz (2011b), Global Crop Yield Reductions due to Surface Ozone Exposure: 2. Year 2030 Potential Crop Production Losses and Economic Damage under Two Scenarios of O_3 Pollution, Atmos. Environ., 45, 2297-2309.

Beer, R., T. A. Glavich, and D. M. Rider (2001), Tropospheric emission spectrometer for the Earth Observing System's Aura satellite, Applied Optics, 40, 2356 – 2367.

Beer, R (2006), TES on the Aura Mission: Scientific Objectives, Measurements, and Analysis Overview, IEEE Transaction on Geoscience and Remote Sensing, 44, 1102-1105.

Bian, J., A. Gettelman, H. Chen, and L. L. Pan (2007), Validation of satellite ozone profile retrievals using Beijing ozonesonde data, J. Geophys. Res., 112, D06305, doi:10.1029/2006JD007502.

Bocquet, M., Elbern, H., Eskes, H., Hirtl, M., Žabkar, R., Carmichael, G. R., Flemming, J., Inness, A., Pagowski, M., Pérez Camaño, J. L., Saide, P. E., San Jose, R., Sofiev, M., Vira, J., Baklanov, A., Carnevale, C., Grell, G., and Seigneur, C. (2015), Data assimilation in atmospheric chemistry models: current status and future prospects for coupled chemistry meteorology models, Atmos. Chem. Phys., 15, 5325-5358, doi:10.5194/acp-15-5325-2015.

Boersma, K. F., Braak, R., van der A, R. J. (2011a), Dutch OMI NO_2 (DOMINO) data product v2.0 HE5 data file user manual. http://www.temis.nl/docs/OMI_NO2_HE5_2.0_2011.pdf.

Boersma, K. F., Eskes, H. J., Dirksen, R. J., van der A, R. J., Veefkind, J. P., Stammes, P., Huijnen, V., Kleipool, Q. L., Sneep, M., Claas, J., Leitão, J., Richter, A., Zhou, Y., Brunner, D. (2011b), An improved tropospheric NO_2 column retrieval algorithm for the Ozone Monitoring Instrument, Atmos. Meas. Tech., 4, 1905-1928.

Bowman, K. W., Rodgers, C. D., Kulawik, S. S., Worden, J., Sarkissian, E., Osterman, G., Steck, T., Lou, M., Eldering, A., Shephard, M., Worden, H., Lampel, M., Clough, S., Brown, P., Rinsland, C., Gunson, M., and Beer, R. (2006), Tropospheric Emission Spectrometer: Retrieval method and error analysis, IEEE Transaction on Geoscience and Remote Sensing, 44 (5), 1297–1307, doi: 10.1109/TGRS.2006.871234.

Bowman, K., and D. K. Henze (2012), Attribution of direct ozone radiative forcing to spatially resolved emissions, Geophys. Res. Lett., 39, L22704, doi:10.1029/2012GL053274.

Brioude, J., Angevine, W. M., Ahmadov, R., Kim, S.-W., Evan, S., McKeen, S. A., Hsie, E.-Y., Frost, G. J., Neuman, J. A., Pollack, I. B., Peischl, J., Ryerson, T. B., Holloway, J., Brown, S. S., Nowak, J. B., Roberts, J. M., Wofsy, S. C., Santoni, G. W., Oda, T., and Trainer, M. (2013), Top-down estimate of surface flux in the Los Angeles Basin using a mesoscale inverse modeling technique: assessing anthropogenic emissions of CO, NO_x and CO_2 and their impacts, Atmos. Chem. Phys., 13, 3661-3677, doi:10.5194/acp-13-3661-2013.

Brown-Steiner, B., and P. Hess (2011), Asian influence on surface ozone in the United States: A comparison of chemistry, seasonality, and transport mechanisms, J. Geophys. Res., 116, D17309, doi:10.1029/2011JD015846.

Cai, C., J. T. Kelly, J. C. Avise, A. P. Kaduwela, and W. R. Stockwell (2011), Photochemical Modeling in California with Two Chemical Mechanisms: Model Intercomparison and Response to Emission Reductions, J. Air & Waste Manage. Assoc., 61:5, 559-572, doi: 10.3155/1047-3289.61.5.559.

Canty, T. P., Hembeck, L., Vinciguerra, T. P., Anderson, D. C., Goldberg, D. L., Carpenter, S. F., Allen, D. J., Loughner, C. P., Salawitch, R. J., and Dickerson, R. R. (2015), Ozone and NOx chemistry in the eastern US: evaluation of CMAQ/CB05 with satellite (OMI) data, Atmos. Chem. Phys., 15, 10965-10982, doi:10.5194/acp-15-10965-2015.

Moved (insertion) [8]

Moved (insertion) [9]

Carmichael, G.R., Tang, Y., Kurata, G., Uno, I., Streets, D.G., Thongboonchoo, N., Woo, J.H., Guttikunda, S., White, A., Wang, T., Blake, D.R., Atlas, E., Fried, A., Potter, B., Avery, M.A., Sachse, G.W., Sandholm, S.T., Kondo, Y., Talbot, R.W., Bandy, A., Thorton, D., and Clarke, A.D. (2003a), Evaluating regional emission estimates using the TRACE-P observations, J. Geophys. Res., 108 (D21), 8810, doi: 10.1029/2002JD003116.

Carmichael, G.R., Tang, Y., Kurata, G., Uno, I., Streets, D., Woo, J.H., Huang, H., Yienger, J., Lefer, B., Shetter, R., Blake, D., Atlas, E., Fried, A., Apel, E., Eisele, F., Cantrell, C., Avery, M., Barrick, J., Sachse, G., Brune, W., Sandholm, S., Kondo, Y., Singh, H., Talbot, R., Bandy, A., Thorton, D., Clarke, A., and Heikes, B. (2003b), Regional-scale chemical transport modeling in support of the analysis of observations obtained during the TRACE-P experiment, J. Geophys. Res., 108 (D21), 8823, doi: 10.1029/2002JD003117.

Carter, W. P. L. (2000), Documentation of the SAPRC-99 chemical mechanism for VOC Reactivity Assessment, final report to California Air Resources Board, Contract No. 92-329 and 95-308.

Cooper, O. R., et al. (2010), Increasing springtime ozone mixing ratios in the free troposphere over western North America, Nature, 463, doi: 10.1038/nature08708.

Cooper, O. R., Oltmans, S. J., Johnson, B. J., Brioude, J., Angevine, W., Trainer, M., Parrish, D. D., Ryerson, T. R., Pollack, I., Cullis, P. D., Ives, M. A., Tarasick, D. W., Al-Saadi, J., and Stajner, I. (2011), Measurement of western U.S. baseline ozone from the surface to the tropopause and assessment of downwind impact regions, J. Geophys. Res., 116, D00V03, doi: 10.1029/2011JD016095.

Cooper, O., et al. (2016), Western NA Performance Evaluation for HTAP2, HTAP2 workshop, Potsdam, Germany, 2016.

Crippa, M., Janssens-Maenhout, G., Dentener, F., Guizzardi, D., Sindelarova, K., Muntean, M., Van Dingenen, R., and Granier, C. (2016), Forty years of improvements in European air quality: regional policy-industry interactions with global impacts, Atmos. Chem. Phys., 16, 3825-3841, doi:10.5194/acp-16-3825-2016.

Emmons, L. K., Hess, P. G., Lamarque, J.-F., and Pfister, G. G. (2012), Tagged ozone mechanism for MOZART-4, CAM-chem and other chemical transport models, Geosci. Model Dev., 5, 1531-1542, doi:10.5194/gmd-5-1531-2012.

Eskes, H. J. and Boersma, K. F. (2003), Averaging kernels for DOAS total-column satellite retrievals, Atmos. Chem. Phys., 3, 1285-1291.

Fiore, A. M., et al. (2009), Multimodel estimates of intercontinental source receptor relationships for ozone pollution, J. Geophys. Res., 114, D04301, doi:10.1029/2008JD010816.

Fiore, A. M., J. T. Oberman, M. Y. Lin, L. Zhang, O. E. Clifton, D. J. Jacob, V. Naik, L. W. Horowitz, J. P. Pinto, and G. P. Milly (2014), Estimating North American background ozone in U.S. surface air with two independent global models: Variability, uncertainties, and recommendations, Atmos. Environ., 96, 284–300, doi: 10.1016/j.atmosenv.2014.07.045.

Flemming, J., Huijnen, V., Arteta, J., Bechtold, P., Beljaars, A., Blechschmidt, A.-M., Diamantakis, M., Engelen, R. J., Gaudel, A., Inness, A., Jones, L., Josse, B., Katragkou, E., Marecal, V., Peuch, V.-H., Richter, A., Schultz, M. G., Stein, O., and Tsikerdekis, A. (2015), Tropospheric chemistry in the Integrated Forecasting System of ECMWF, Geosci. Model Dev., 8, 975-1003, doi:10.5194/gmd-8-975-2015.

Flemming, J., Benedetti, A., Inness, A., Engelen, R., Jones, L., Huijnen, V., Remy, S., Parrington, M., Suttie, M., Bozzo, A., Peuch, V.-H., Akritidis, D., and Katragkou, E. (2017), The CAMS

Moved (insertion) [10]

Formatted: Default Paragraph Font, Font color: Black

Moved (insertion) [11]

Formatted: Default Paragraph Font, Font color: Black

Deleted: 6

interim Reanalysis of Carbon Monoxide, Ozone and Aerosol for 2003–2015, Atmos. Chem. Phys., 17, 1945-1983, doi:10.5194/acp-17-1945-2017.

Galmarini, S., C. Hogrefe, D. Brunner, P. Makar, A. Baklanov (2015), Preface to the AQMEII p2 Special issue, Atmos. Environ., 115, 340-344.

Galmarini, S., Koffi, B., Solazzo, E., Keating, T., Hogrefe, C., Schulz, M., Benedictow, A., Griesfeller, J. J., Janssens-Maenhout, G., Carmichael, G., Fu, J., and Dentener, F. (2017), Technical note: Coordination and harmonization of the multi-scale, multi-model activities HTAP2, AQMEII3, and MICS-Asia3: simulations, emission inventories, boundary conditions, and model output formats, Atmos. Chem. Phys., 17, 1543-1555, doi:10.5194/acp-17-1543-2017.

Geddes, J. A., Heald, C. L., Silva, S. J., and Martin, R. V. (2016), Land cover change impacts on atmospheric chemistry: simulating projected large-scale tree mortality in the United States, Atmos. Chem. Phys., 16, 2323-2340, doi:10.5194/acp-16-2323-2016.

Gery, M. W., G. Z. Whitten, J. P. Killus, and M. C. Dodge (1989), A photochemical kinetics mechanism for urban and regional scale computer modeling, J. Geophys. Res., 94, 12,925 – 12,956, doi:10.1029/JD094iD10p12925.

Granier, C., Lamarque, J. F., Mieville, A., Muller, J. F., Olivier, J., Orlando, J., Peters, J., Petron, G., Tyndall, G., and Wallens, S. (2005), POET, a database of surface emissions of ozone precursors, http://www.aero.jussieu.fr/projet/ACCENT/POET.php.

Gratz, L.E., Jaffe, D.A., and Hee, J.R. (2014), Causes of increasing ozone and decreasing carbon monoxide in springtime at the Mt. Bachelor Observatory from 2004 to 2013, Atmos. Environ., 109, 323–330, doi: 10.1016/j.atmosenv.2014.05.076.

Guenther, A. B., X. Jiang, C. L. Heald, T. Sakulyanontvittaya, T. Duhl, L. K. Emmons, and X. Wang (2012), The Model of Emissions of Gases and Aerosols from Nature version 2.1 (MEGAN2.1): an extended and updated framework for modeling biogenic emissions, Geosci. Model Dev., 5 (6), 1471-1492.

Henze, D. K., Hakami, A., and Seinfeld, J. H. (2007), Development of the adjoint of GEOS-Chem, Atmos. Chem. Phys., 7, 2413–2433, doi:10.5194/acp-7-2413-2007.

Hilsenrath, E., and K. Chance (2013), NASA ups the TEMPO on monitoring air pollution, Earth Obs., 25, 10–15.

Hogrefe, C., Isukapalli, S., Tang, X., Georgopoulos, P., He, S., Zalewsky, E., Hao, W., Ku, J., Key, T., and Sistla, G. (2011), Impact of biogenic emission uncertainties on the simulated response of ozone and fine Particulate Matter to anthropogenic emission reductions, J. Air Waste Manage., 61, 92–108.

Huang, M., Carmichael, G. R., Adhikary, B., Spak, S. N., Kulkarni, S., Cheng, Y. F., Wei, C., Tang, Y., Parrish, D. D., Oltmans, S. J., D'Allura, A., Kaduwela, A., Cai, C., Weinheimer, A. J., Wong, M., Pierce, R. B., Al-Saadi, J. A., Streets, D. G., and Zhang, Q. (2010), Impacts of transported background ozone on California air quality during the ARCTAS-CARB period – a multi-scale modeling study, Atmos. Chem. Phys., 10, 6947-6968, doi: 10.5194/acp-10-6947-2010.

Huang, M., Carmichael, G. R., Chai, T., Pierce, R. B., Oltmans, S. J., Jaffe, D. A., Bowman, K. W., Kaduwela, A., Cai, C., Spak, S. N., Weinheimer, A. J., Huey, L. G., and Diskin, G. S. (2013a), Impacts of transported background pollutants on summertime western US air quality: model evaluation, sensitivity analysis and data assimilation, Atmos. Chem. Phys., 13, 359-391, doi: 10.5194/acp-13-359-2013.

Deleted: . Discuss.,

Deleted: 2016-666, in review

Deleted: 6

Deleted: Harmonization

Deleted: HTAP, AQMEII

Deleted: . Discuss.,

Deleted: 2016-828, in review.

Deleted: Grewe, V., Dahlmann, K., Matthes, S., and Steinbrecht, W. (2012), Attributing ozone to NOx emissions: Implications for climate mitigation measures, Atmos.

Moved down [12]: Environ.,

Formatted: Default Paragraph Font, Font:(Asian) +Theme Body Asian (DengXian), (Asian) Chinese (PRC)

Deleted: 59, 102-107, doi: 10.1016/j.atmosenv.2012.05.002.

Huang, M., Bowman, K. W., Carmichael, G. R., Pierce, R. B., Worden, H. M., Luo, M., Cooper, O. R., Pollack, I. B., Ryerson, T. B., Brown, S. S. (2013b), Impact of southern California anthropogenic emissions on ozone pollution in the mountain states, J. Geophys. Res., 118, 12784-12803, doi: 10.1002/2013JD020205.

Huang, M., et al. (2014), Changes in nitrogen oxides emissions in California during 2005–2010 indicated from top-down and bottom-up emission estimates, J. Geophys. Res., 119, 12,928–12,952, doi: 10.1002/2014JD022268, 2014.

Huang, M., et al. (2015), Improved Western US Background Ozone Estimates via Constraining Nonlocal and Local Source Contributions using Aura TES and OMI Observations, J. Geophys. Res., 120, 3572–3592, doi: 10.1002/2014JD022993.

Huang, M., Carmichael, G. R., Crawford, J. H., Wisthaler, A., Zhan, X., Hain, C. R., Lee, P., and Guenther, A. B. (2017), Linkages between land initialization of the NASA-Unified WRF v7 and biogenic isoprene emission estimates during the SEAC4RS and DISCOVER-AQ airborne campaigns, Geosci. Model Dev. Discuss., doi:10.5194/gmd-2017-13, in review.

Inness, A., Blechschmidt, A.-M., Bouarar, I., Chabrillat, S., Crepulja, M., Engelen, R. J., Eskes, H., Flemming, J., Gaudel, A., Hendrick, F., Huijnen, V., Jones, L., Kapsomenakis, J., Katragkou, E., Keppens, A., Langerock, B., de Mazière, M., Melas, D., Parrington, M., Peuch, V. H., Razinger, M., Richter, A., Schultz, M. G., Suttie, M., Thouret, V., Vrekoussis, M., Wagner, A., and Zerefos, C. (2015), Data assimilation of satellite-retrieved ozone, carbon monoxide and nitrogen dioxide with ECMWF's Composition-IFS, Atmos. Chem. Phys., 15, 5275-5303, doi:10.5194/acp-15-5275-2015.

Jaffe, D.A. (2011), Relationship between surface and free tropospheric ozone in the Western U.S., Environ. Sci. Technol., 45, 432–438, doi: 10.1021/es1028102.

Janssens-Maenhout, G., Crippa, M., Guizzardi, D., Dentener, F., Muntean, M., Pouliot, G., Keating, T., Zhang, Q., Kurokawa, J., Wankmüller, R., Denier van der Gon, H., Kuenen, J. J. P., Klimont, Z., Frost, G., Darras, S., Koffi, B., and Li, M. (2015), HTAP_v2.2: a mosaic of regional and global emission grid maps for 2008 and 2010 to study hemispheric transport of air pollution, Atmos. Chem. Phys., 15, 11411-11432, doi:10.5194/acp-15-11411-2015.

Jacob, D. J., Logan, J. A., and Murti, P. P. (1999), Effect of rising Asian emissions on surface ozone in the United States, Geophys. Res. Lett., 26, 2175-2178, doi: 10.1029/1999GL900450.

Jerret, M., R. T. Burnett, C. A. Popo, III, K. Ito, G. Thurston, D. Krewski, Y. Shi, E. Calle, and M. Thun (2009), Long-Term Ozone Exposure and Mortality, the New England Journal of Medicine, 360, 1085-1096, doi: 10.1056/NEJMoa0803894.

Kaiser, J. W., Heil, A., Andreae, M. O., Benedetti, A., Chubarova, N., Jones, L., Morcrette, J.-J., Razinger, M., Schultz, M. G., Suttie, M., and van der Werf, G. R. (2012), Biomass burning emissions estimated with a global fire assimilation system based on observed fire radiative power, Biogeosciences, 9, 527–554, doi:10.5194/bg-9-527-2012.

Kalnay, E., and Co-authors (1996), The NCEP/NCAR 40-Year Reanalysis Project, Bulletin of the American Meteorological Society, 77, 437–471.

Kim, S.-W., B. C. McDonald, S. Baidar, S. S. Brown, B. Dube, R. A. Ferrare, G. J. Frost, R. A. Harley, J. S. Holloway, H.-J. Lee, et al. (2016), Modeling the weekly cycle of NO_x and CO emissions and their impacts on O_3 in the Los Angeles-South Coast Air Basin during the CalNex 2010 field campaign, J. Geophys. Res. Atmos., 121, 1340–1360, doi:10.1002/2015JD024292.

Koffi, B., Dentener, F., Janssens-Maenhout, G., Guizzardi, D., Crippa, M., Diehl, T., Galmarini, S., and Solazzo, E., Hemispheric Transport Air Pollution (HTAP): Specification of the HTAP2

Deleted: F.

Deleted: G.

Deleted: D.

Deleted: M.

Deleted: T.

Deleted: S.

Deleted: E.

Deleted: (2016),

experiments – Ensuring harmonized modelling, EUR 28255 EN – Scientific and Technical Research Reports, doi:10.2788/725244, 2016.

Langford, A. O., Brioude, J., Cooper, O.R., Senff, C.J., Alvarez II, R.J., Hardesty, R.M., Johnson, B.J., and Oltmans, S.J. (2011), Stratospheric influence on surface ozone in the Los Angeles area during late spring and early summer of 2010, J. Geophys. Res. Atmos., 117, D00V06, doi: 10.1029/2011JD016766.

Lapina, K., D. K. Henze, J. B. Milford, M. Huang, M. Lin, A. M. Fiore, G. Carmichael, G. G. Pfister, and K. Bowman (2014), Assessment of source contributions to seasonal vegetative exposure to ozone in the U.S., J. Geophys. Res. Atmos., 119, 324–340, doi:10.1002/2013JD020905.

Levelt, P.F., E. Hilsenrath, G.W. Leppelmeier, G.H.J. van den Oord, P.K. Bhartia, J. Tamminen, J.F. de Haan and J.P. Veefkind (2006), Science Objectives of the Ozone Monitoring Instrument, IEEE Transaction on Geoscience and Remote Sensing, 44, 1199-1208.

Li, M., Zhang, Q., Kurokawa, J.-I., Woo, J.-H., He, K., Lu, Z., Ohara, T., Song, Y., Streets, D. G., Carmichael, G. R., Cheng, Y., Hong, C., Huo, H., Jiang, X., Kang, S., Liu, F., Su, H., and Zheng, B. (2017), MIX: a mosaic Asian anthropogenic emission inventory under the international collaboration framework of the MICS-Asia and HTAP, Atmos. Chem. Phys., 17, 935-963, doi:10.5194/acp-17-935-2017.

Lin, M., Holloway, T., Carmichael, G. R., and Fiore, A. M. (2010), Quantifying pollution inflow and outflow over East Asia in spring with regional and global models, Atmos. Chem. Phys., 10, 4221-4239, doi:10.5194/acp-10-4221-2010.

Lin, M., A. M. Fiore, L. W. Horowitz, O. R. Cooper, V. Naik, J. Holloway, B. J. Johnson, A. Middlebrook, S. J. Oltmans, I. B. Pollack, T. B. Ryerson, J. X. Warner, C. Wiedinmyer, J. Wilson, B. Wyman (2012a), Transport of Asian ozone pollution into surface air over the western United States in spring, J. Geophys. Res., 117, D00V07, doi: 10.1029/2011JD016961.

Lin, M., A. Fiore, O. R. R. Cooper, L. W. Horowitz, A. O. O. Langford, H. Levy II, B. J. Johnson, V. Naik, S. J. Oltmans, and C. J. Senff (2012b), Springtime high surface ozone events over the western United States: Quantifying the role of stratospheric intrusions, J. Geophys. Res., 117, D00V22, doi: 10.1029/2012JD018151.

Lin, M., L.W. Horowitz, S. J. Oltmans, A. M. Fiore, S. Fan (2014), Tropospheric ozone trends at Manna Loa Observatory tied to decadal climate variability, Nature Geoscience, 7, 136-143, doi:10.1038/NGEO2066.

Lin, M., L. W. Horowitz, O. R. Cooper, D. Tarasick, S. Conley, L. T. Iraci, B. Johnson, T. Leblanc, I. Petropavlovskikh, and E. L. Yates (2015), Revisiting the evidence of increasing springtime ozone mixing ratios in the free troposphere over western North America, Geophys. Res. Lett., 42, 8719–8728, doi:10.1002/2015GL065311.

Lin, M., Horowitz, L. W., Payton, R., Fiore, A. M., and Tonnesen, G. (2016), US surface ozone trends and extremes from 1980–2014: Quantifying the roles of rising Asian emissions, domestic controls, wildfires, and climate, Atmos. Chem. Phys. Discuss., doi:10.5194/acp-2016-1093, in review.

Liu, F., Q. Zhang, R. J. van der A, B. Zheng, D. Tong, L. Yan, Y. Zheng, and K. He (2016), Recent reduction in NO_x emissions over China: Synthesis of satellite observations and emission inventories, Environ. Res. Lett., 11 (11), 114002, doi: 10.1088/1748-9326/11/11/114002.

Livesey, N.J., M.J. Filipiak, L. Froidevaux, W.G. Read, A. Lambert, M.L. Santee, J.H. Jiang, H.C. Pumphrey, J.W. Waters, R.E. Cofield, D.T. Cuddy, W.H. Daffer, B.J. Drouin, R.A. Fuller, R.F. Jarnot, Y.B. Jiang, B.W. Knosp, Q.B. Li, V.S. Perun, M.J. Schwartz, W.V. Snyder, P.C. Stek,

Deleted: -

Deleted: in preparation.

Deleted: . B

Deleted: . F

Deleted: . P

Deleted: . J

Deleted: . C

Deleted: 5

Deleted: for

Deleted: the

Deleted: projects

Deleted: . Discuss., 15, 34813-34869

Deleted: acpd-15-34813-2015

Moved (insertion) [13]

R.P. Thurstans, P.A. Wagner, M. Avery, E.V. Browell, J-P. Cammas, L.E. Christensen, G.S. Diskin, R-S. Gao, H-J. Jost, M. Loewenstein, J.D. Lopez, P. Nedelec, G.B. Osterman, G.W. Sachse, and C.R. Webster (2008), Validation of Aura Microwave Limb Sounder O3 and CO observations in the upper troposphere and lower stratosphere, J. Geophys. Res. 113, D15S02, doi:10.1029/2007JD008805.

Luecken, D.J., S. Phillips, G. Sarwar, C. Jang, Effects of using the CB05 vs. SAPRC99 vs. CB4 chemical mechanism on model predictions (2008), Ozone and gas-phase photochemical precursor concentrations, Atmos. Environ., 42 (23), 5805-5820, doi: 10.1016/j.atmosenv.2007.08.056.

Maas, R. and P. Grennfelt (eds) (2016), Towards Cleaner Air Scientific Assessment Report 2016. EMEP Steering Body and Working Group on Effects of the Convention on Long-Range Transboundary Air Pollution, Oslo, http://www.unece.org/fileadmin/DAM/env/lrtap/ExecutiveBody/35th_session/CLRTAP_Scientific_Assessment_Report_-_Final_20-5-2016.pdf.

Madronich, S., Flocke, S., Zeng, J., Petropavlovskikh, I., and Lee-Taylor, J. (2002), The Tropospheric Ultra-violet Visible (TUV) model Manual, https://www2.acom.ucar.edu/modeling/tropospheric-ultraviolet-and-visible-tuv-radiation-model.

Mauzerall, D. L. and Wang, X. (2001), Protecting Agricultural Crops from the Effects of Tropospheric Ozone Exposure: Reconciling Science and Standard Setting in the United States, Europe and Asia, Annual Review of Energy and the Environment, 26, 237-268.

McDonald-Buller, E. C., et al. (2011), Establishing policy relevant background (PRB) ozone concentrations in the United States, Environ. Sci. Technol., 45, 9484–9497.

Meijer, E. W., van Velthoven, P. F. J., Brunner, D. W., Huntrieser, H., and Kelder, H. (2001), Improvement and evaluation of the parameterization of nitrogen oxide production by lightning, Phys. Chem. Earth Pt. C, 26, 577–583.

Mesinger, F., DiMego, G., Kalnay, E., Mitchell, K., Shafran, P. C., Ebisuzaki, W., Jovic, D., Woollen, J., Rogers, E., Berbery, E. H., Ek, M. B., Fan, Y., Grumbine, R., Higgins, W., Li, H., Lin, Y., Manikin, G., Parrish, D. and Shi, W. (2006), North American Regional Reanalysis, Bulletin of the American Meteorological Society, 87(3), 343–360, doi: 10.1175/BAMS-87-3-343.

Miyazaki, K., Eskes, H. J., Sudo, K., Takigawa, M., van Weele, M., Boersma, K. F. (2012), Simultaneous assimilation of satellite NO_2, O_3, CO, and HNO_3 data for the analysis of tropospheric chemical composition and emissions, Atmos. Chem. Phys., 12, 9545-9579.

Monks, P. S., Archibald, A. T., Colette, A., Cooper, O., Coyle, M., Derwent, R., Fowler, D., Granier, C., Law, K. S., Mills, G. E., Stevenson, D. S., Tarasova, O., Thouret, V., von Schneidemesser, E., Sommariva, R., Wild, O., and Williams, M. L. (2015), Tropospheric ozone and its precursors from the urban to the global scale from air quality to short-lived climate forcer, Atmos. Chem. Phys., 15, 8889-8973, doi:10.5194/acp-15-8889-2015.

Murray, L. T., D. J. Jacob, J. A. Logan, R. C. Hudman, and W. J. Koshak (2012), Optimized regional and interannual variability of lightning in a global chemical transport model constrained by LIS/OTD satellite data, J. Geophys. Res., 117, D20307, doi:10.1029/2012JD017934.

Moved up [13]: Liu, F., Q. Zhang, R. J. van der A, B. Zheng, D. Tong, L. Yan, Y. Zheng, and K. He (2016), Recent reduction in NO_x emissions over China: Synthesis of satellite observations and emission inventories, Environ. Res. Lett., 11 (11), 114002, doi: 10.1088/1748-9326/11/11/114002.

National Research Council (NRC) (2009), global sources of local pollution-An Assessment of Long-Range Transport of Key Air Pollutants to and from the United States, 35-66, http://books.nap.edu/openbook.php?record_id=12743&page=35.

Neuman, J. A., et al. (2012), Observations of ozone transport from the free troposphere to the Los Angeles basin, J. Geophys. Res. Atmos., 117, D00V09, doi: 10.1029/2011JD016919.

Oetjen, H., Payne, V. H., Kulawik, S. S., Eldering, A., Worden, J., Edwards, D. P., Francis, G. L., Worden, H. M., Clerbaux, C., Hadji-Lazaro, J., and Hurtmans, D. (2014), Extending the satellite data record of tropospheric ozone profiles from Aura-TES to MetOp-IASI: characterisation of optimal estimation retrievals, Atmos. Meas. Tech., 7, 4223–4236, doi:10.5194/amt-7-4223- 2014.

Oetjen, H., Payne, V. H., Neu, J. L., Kulawik, S. S., Edwards, D. P., Eldering, A., Worden, H. M., and Worden, J. R. (2016), A joint data record of tropospheric ozone from Aura-TES and MetOp-IASI, Atmos. Chem. Phys., 16, 10229-10239, doi:10.5194/acp-16-10229-2016.

Ott, L. E., B. N. Duncan, A. M. Thompson, G. Diskin, Z. Fasnacht, A. O. Langford, M. Lin, A. M. Molod, J. E. Nielsen, S. E. Pusede, et al. (2016), Frequency and impact of summertime stratospheric intrusions over Maryland during DISCOVER-AQ (2011): New evidence from NASA's GEOS-5 simulations, J. Geophys. Res. Atmos., 121, 3687–3706, doi:10.1002/2015JD024052.

Park, R. J., D. J. Jacob, B. D. Field, R. M. Yantosca, and M. Chin (2004), Natural and transboundary pollution influences on sulfate-nitrate-ammonium aerosols in the United States: Implications for policy, J. Geophys. Res., 109, D15204, doi:10.1029/2003JD004473.

Parrington, M., D. B. A. Jones, K. W. Bowman, L. W. Horowitz, A. M. Thompson, D. W. Tarasick, and J. C. Witte (2008), Estimating the summertime tropospheric ozone distribution over North America through assimilation of observations from the Tropospheric Emission Spectrometer, J. Geophys. Res., 113, D18307, doi:10.1029/2007JD009341.

Parrington, M., D. B. A. Jones, K. W. Bowman, A. M. Thompson, D. W. Tarasick, J. Merrill, S. J. Oltmans, T. Leblanc, J. C. Witte, and D. B. Millet (2009), Impact of the assimilation of ozone from the Tropospheric Emission Spectrometer on surface ozone across North America, Geophys. Res. Lett., 36, L04802, doi:10.1029/2008GL036935.

Parrish, D. D., D. B. Millet, and A. H. Goldstein (2009), Increasing ozone in marine boundary layer inflow at the west coasts of North America and Europe, Atmos. Chem. Phys., 9, 1303–1323, doi:10.5194/acp-9-1303-2009.

Parrish, D. D., Aikin, K. C., Oltmans, S. J., Johnson, B. J., Ives, M., and Sweeny, C. (2010), Impact of transported background ozone inflow on summertime air quality in a California ozone exceedance area, Atmos. Chem. Phys., 10, 10093-10109, doi:10.5194/acp-10-10093-2010.

Parrish, D. D., et al. (2012), Long-term changes in lower tropospheric baseline ozone concentrations at northern mid-latitudes, Atmos. Chem. Phys., 12, 11,485–11,504, doi:10.5194/acp-12-11485-2012.

Pierce, R. B., et al. (2007), Chemical data assimilation estimates of continental U.S. ozone and nitrogen budgets during the Intercontinental Chemical Transport Experiment–North America, J. Geophys. Res., 112, D12S21, doi:10.1029/2006JD007722.

Pierce, R. B., et al. (2009), Impacts of background ozone production on Houston and Dallas, Texas, air quality during the Second Texas Air Quality Study field mission, J. Geophys. Res., 114, D00F09, doi:10.1029/2008JD011337.

Pouliot, G., H. A.C. Denier van der Gon, J. Kuenen, J. Zhang, M. D. Moran, P.A. Makar (2015), Analysis of the emission inventories and model-ready emission datasets of Europe and North America for phase 2 of the AQMEII project, Atmos. Environ., 115, 345-360.

Qu, Z., D. K. Henze, S. L. Capps, Y. Wang, X. Xu, J. Wang (2016), Monthly top-down NO_x emissions for China (2005-2012): a hybrid inversion method and trend analysis, submitted.

Quennehen, B., Raut, J.-C., Law, K. S., Daskalakis, N., Ancellet, G., Clerbaux, C., Kim, S.-W., Lund, M. T., Myhre, G., Olivié, D. J. L., Safieddine, S., Skeie, R. B., Thomas, J. L., Tsyro, S., Bazureau, A., Bellouin, N., Hu, M., Kanakidou, M., Klimont, Z., Kupiainen, K., Myriokefalitakis, S., Quaas, J., Rumbold, S. T., Schulz, M., Cherian, R., Shimizu, A., Wang, J., Yoon, S.-C., and Zhu, T. (2016), Multi-model evaluation of short-lived pollutant distributions over east Asia during summer 2008, Atmos. Chem. Phys., 16, 10765-10792, doi:10.5194/acp-16-10765-2016.

Reidmiller, D. R., Fiore, A. M., Jaffe, D. A., Bergmann, D., Cuvelier, C., Dentener, F. J., Duncan, B. N., Folberth, G., Gauss, M., Gong, S., Hess, P., Jonson, J. E., Keating, T., Lupu, A., Marmer, E., Park, R., Schultz, M. G., Shindell, D. T., Szopa, S., Vivanco, M. G., Wild, O., and Zuber, A. (2009), The influence of foreign vs. North American emissions on surface ozone in the US, Atmos. Chem. Phys., 9, 5027-5042, doi:10.5194/acp-9-5027-2009.

Rodgers, C. D. (2000), Inverse Methods for Atmospheric Sounding: Theory and Practice, World Sci., Singapore.

Ryerson, T. B., Andrews, A. E., Angevine, W. M., Bates, T. S., Brock, C. A., Cairns, B., Cohen, R. C., Cooper, O. R., de Gouw, J. A., Fehsenfeld, F. C., Ferrare, R. A., Fischer, M. L., Flagan, R. C., Goldstein, A. H., Hair, J. W., Hardesty, R. M., Hostetler, C. A., Jimenez, J. L., Langford, A. O., McCauley, E., McKeen, S. A., Molina, L. T., Nenes, A., Oltmans, S. J., Parrish, D. D., Pederson, J. R., Pierce, R. B., Prather, K., Quinn, P. K., Seinfeld, J. H., Senff, C. J., Sorooshian, A., Stutz, J., Surratt, J. D., Trainer, M., Volkamer, R., Williams, E. J., Wofsy, S. C. (2013), The 2010 California Research at the Nexus of Air Quality and Climate Change (CalNex) field study, J. Geophys. Res., 118, 5830–5866.

Schere, K. J. Flemming, R. Vautard, C. Chemel, A. Colette, C. Hogrefe, B. Bessagnet, F. Meleux, R. Mathur, S. Roselle, R.-M. Hu, R. S. Sokhi, S. T. Rao, S. Galmarini (2012), Trace gas/aerosol boundary concentrations and their impacts on continental-scale AQMEII modeling domains, Atmos. Environ., 53, 38-50, doi: 10.1016/j.atmosenv.2011.09.043.

Shindell, D. T., G. Faluvegi, D. M. Koch, G. A. Schmidt, N. Unger, and S. E. Bauer (2009), Improved attribution of climate forcing to emissions, Science, 326, 716–718, doi: 10.1126/science.1174760.

Shindell, D. T., et al. (2013), Radiative forcing in the ACCMIP historical and future climate simulations, Atmos. Chem. Phys., 13, 2939–2974, doi:10.5194/acp-13-2939-2013.

Simpson, D., Benedictow, A., Berge, H., Bergström, R., Emberson, L. D., Fagerli, H., Flechard, C. R., Hayman, G. D., Gauss, M., Jonson, J. E., Jenkin, M. E., Nyíri, A., Richter, C., Semeena, V. S., Tsyro, S., Tuovinen, J.-P., Valdebenito, A., and Wind, P. (2012), The EMEP MSC-W chemical transport model – technical description, Atmos. Chem. Phys., 12, 7825–7865, doi:10.5194/acp-12- 7825-2012.

Sindelarova, K., Granier, C., Bouarar, I., Guenther, A., Tilmes, S., Stavrakou, T., Müller, J.-F., Kuhn, U., Stefani, P., and Knorr, W. (2014), Global data set of biogenic VOC emissions calculated by the MEGAN model over the last 30 years, Atmos. Chem. Phys., 14, 9317–9341, doi:10.5194/acp-14-9317-2014.

Moved (insertion) [14]

Formatted: apple-style-span, Font:Arial, Font color: Black, Pattern: Clear (White)

Skamarock, W. C., J. B. Klemp, J. Dudhia, D. Gill, D. M. Barker, W. Wang, and J. G. Powers (2008), A description of the Advanced Research WRF version 3 (Available at www.mmm.ucar.edu/wrf/users/docs/arwv3.pdf).

Smith, K. R., Jerrett, M., and Anderson, H. R. et al. (2009), Public health benefits of strategies to reduce greenhouse-gas emissions: health implications of short-lived greenhouse pollutants, Lancet, doi: 10.1016/S0140-6736 (09) 61716-5.

Solazzo, E. R. Bianconi, R. Vautard, K. W. Appel, M. D. Moran, C. Hogrefe, B. Bessagnet, J. Brandt, J. H. Christensen, C. Chemel, I. Coll, H. D. van der Gon, J. Ferreira, R. Forkel, X. V. Francis, G. Grell, P. Grossi, A. B. Hansen, A. Jeričević, L. Kraljević, A. I. Miranda, U. Nopmongcol, G. Pirovano, M. Prank, A. Riccio, K. N. Sartelet, M. Schaap, J. D. Silver, R. S. Sokhi, J. Vira, J. Werhahn, R. Wolke, G. Yarwood, J. Zhang, S.T. Rao, S. Galmarini (2012), Model evaluation and ensemble modelling of surface-level ozone in Europe and North America in the context of AQMEII, Atmos. Environ., 53, 60-74, , doi: 10.1016/j.atmosenv.2012.01.003.

Søvde, O. A., Prather, M. J., Isaksen, I. S. A., Berntsen, T. K., Stordal, F., Zhu, X., Holmes, C. D., and Hsu, J. (2012), The chemical transport model Oslo CTM3, Geosci. Model Dev., 5, 1441–1469, doi:10.5194/gmd-5-1441-2012.

Sudo, K., M. Takahashi, J. Kurokawa, and H. Akimoto (2002), Chaser: A global chemical model of the troposphere, 1. Model description, J. Geophys. Res., 107(D17), 4339, doi:10.1029/2001JD001113.

Stevenson, D. S., et al. (2006), Multimodel ensemble simulations of present-day and near-future tropospheric ozone, J. Geophys. Res., 111, D08301, doi:10.1029/2005JD006338.

Stevenson, D. S., et al. (2013), Tropospheric ozone changes, radiative forcing and attribution to emissions in the Atmospheric Chemistry and Climate Model Intercomparison Project (ACCMIP), Atmos. Chem. Phys., 13, 3063–3085, doi:10.5194/acp-13-3063-2013.

Susaya, J., Kim, K.-H., Shon, Z.-H., Brown R. J. (2013), Demonstration of long-term increases in tropospheric O_3 levels: Causes and potential impacts, Chemosphere, 92, 1520–1528.

Task Force on Hemispheric Transport of Air Pollution (HTAP) (2010), 2010 Final Assessment report, Part A: Ozone and particulate matter, http://www.htap.org/activities/2010_Final_Report/HTAP%202010%20Part%20A%2011040 7.pdf.

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. (2016), Representation of the Community Earth System Model (CESM1) CAM4-chem within the Chemistry- Climate Model Initiative (CCMI), Geosci. Model Dev., 9, 1853– 1890, doi:10.5194/gmd-9-1853-2016.

Travis, K. R., Jacob, D. J., Fisher, J. A., Kim, P. S., Marais, E. A., Zhu, L., Yu, K., Miller, C. C., Yantosca, R. M., Sulprizio, M. P., Thompson, A. M., Wennberg, P. O., Crounse, J. D., St. Clair, J. M., Cohen, R. C., Laugher, J. L., Dibb, J. E., Hall, S. R., Ullmann, K., Wolfe, G. M., Pollack, I. B., Peischl, J., Neuman, J. A., and Zhou, X. (2016), Why do models overestimate surface ozone in the Southeast United States?, Atmos. Chem. Phys., 16, 13561-13577, doi:10.5194/acp-16-13561-2016.

United Nations Environment Programme and World Meteorological Organization (2011), Integrated Assessment of Black Carbon and Tropospheric Ozone: Summary for Decision Makers, http://www.unep.org/dewa/Portals/67/pdf/Black_Carbon.pdf.

Moved (insertion) [15]

Moved up [10]: H.,

Deleted: Stjern, C.

Moved up [8]: W.,

Deleted: Samset, B.

Moved up [11]: Flemming, J.,

Deleted: Phys., 16, 13579-13599, doi:10.5194/acp-16-13579-2016.

Formatted: Default Paragraph Font, Font color: Black

Deleted: Haslerud, A. S., Henze, D

Moved up [14]: ., Jonson, J.

Formatted: Default Paragraph Font, Font color: Black

Moved up [15]: Sudo, K.,

Moved up [9]: Chem.

Deleted: Myhre, G., Bian, H., Chin, M., Davila, Y., Dentener, F., Emmons, L.,

Formatted: apple-style-span, Font color: Black, Pattern: Clear (White)

Deleted: E., Kucsera, T., Lund, M. T., Schulz, M.,

Deleted: Takemura, T., and Tilmes, S. (2016), Atmos.

US EPA (2016a), Implementation of the 2015 Primary Ozone NAAQS: Issues Associated with Background Ozone White Paper for Discussion, https://www.epa.gov/sites/production/files/2016-03/documents/whitepaper-bgo3-final.pdf.

US EPA (2016b), High level summary of background ozone workshop, https://www.epa.gov/sites/production/files/2016-03/documents/bgo3-high-level-summary.pdf.

van der Werf, G. R., Randerson, J. T., Giglio, L., Collatz, G. J., Mu, M., Kasibhatla, P. S., Morton, D. C., DeFries, R. S., Jin, Y., and van Leeuwen, T. T. (2010), Global fire emissions and the contribution of deforestation, savanna, forest, agricultural, and peat fires (1997–2009), Atmos. Chem. Phys., 10, 11707-11735, doi:10.5194/acp-10-11707-2010.

van Noije, T. P. C., Eskes, H. J., Dentener, F. J., Stevenson, D. S., Ellingsen, K., Schultz, M. G., Wild, O., Amann, M., Atherton, C. S., Bergmann, D. J., Bey, I., Boersma, K. F., Butler, T., Cofala, J., Drevet, J., Fiore, A. M., Gauss, M., Hauglustaine, D. A., Horowitz, L. W., Isaksen, I. S. A., Krol, M. C., Lamarque, J.-F., Lawrence, M. G., Martin, R. V., Montanaro, V., Müller, J.-F., Pitari, G., Prather, M. J., Pyle, J. A., Richter, A., Rodriguez, J. M., Savage, N. H., Strahan, S. E., Sudo, K., Szopa, S., and van Roozendael, M. (2006), Multi-model ensemble simulations of tropospheric NO_2 compared with GOME retrievals for the year 2000, Atmos. Chem. Phys., 6, 2943-2979, doi:10.5194/acp-6-2943-2006.

Verstraeten, W. W., K. F. Boersma, J. Zörner, M. A. F. Allaart, K. W. Bowman, and J. R. Worden (2013), Validation of six years of TES tropospheric ozone retrievals with ozonesonde measurements: Implications for spatial patterns and temporal stability in the bias, Atmos. Meas. Tech., 6, 1413–1423.

Verstraeten, W.W., J. L. Neu, J. E. Williams, K. W. Bowman, J. R. Worden, and K. F. Boersma (2015), Rapid increases in tropospheric ozone production and export from China, Nature Geoscience, 8, 690–695, doi:10.1038/ngeo2493.

Wang, H., D. J. Jacob, P. L. Sager, D. G. Streets, R. J. Park, A. B. Gilliland, and A. van Donkelaar (2009), Surface ozone background in the United States: Canadian and Mexican pollution influences, Atmos. Environ., 43(6), 1310–1319, doi:10.1016/j.atmosenv.2008.11.036.

Wang, Y., Konopka, P., Liu, Y., Chen, H., Müller, R., Plöger, F., Riese, M., Cai, Z., and Lü, D. (2012), Tropospheric ozone trend over Beijing from 2002–2010: ozonesonde measurements and modeling analysis, Atmos. Chem. Phys., 12, 8389-8399, doi:10.5194/acp-12-8389-2012.

Warneke, C., J. A. deGouw, J. S. Holloway, J. Peischl, T. B. Ryerson, E. Atlas, D. Blake, M. Trainer, and D. D. Parrish (2012), Multiyear trends in volatile organic compounds in Los Angeles, California: Five decades of decreasing emissions, J. Geophys. Res., 117, D00V17, doi:10.1029/2012JD017899.

Warner, J. X., McCourt Comer, M., Barnet, C. D., McMillan, W. W., Wolf, W., Maddy, E., and Sachse, G. (2007), A comparison of satellite tropospheric carbon monoxide measurements from AIRS and MOPITT during INTEX-A, J. Geophys. Res., 112, D12S17, doi:10.1029/2006JD007925, 2007.

Wiedinmyer, C., Akagi, S. K., Yokelson, R. J., Emmons, L. K., Al-Saadi, J. A., Orlando, J. J., and Soja, A. J. (2011), The Fire INventory from NCAR (FINN): a high resolution global model to estimate the emissions from open burning, Geosci. Model Dev., 4, 625-641, doi:10.5194/gmd-4-625-2011.

Wigder, N.L., Jaffe, D.A., Herron-Thorpe, F.L., and Vaughan, J.K. (2013), Influence of daily variations in baseline ozone on urban air quality in the United States Pacific Northwest, J. Geophys. Res., 118, 3343–3354, doi: 10.1029/2012JD018738.

Formatted: Default Paragraph Font, Font:Arial, Font color: Auto, Pattern: Clear

Moved (insertion) [12]

Formatted: Default Paragraph Font, Font:Arial

Wild, O., Fiore, A. M., Shindell, D. T., Doherty, R. M., Collins, W. J., Dentener, F. J., Schultz, M. G., Gong, S., MacKenzie, I. A., Zeng, G., Hess, P., Duncan, B. N., Bergmann, D. J., Szopa, S., Jonson, J. E., Keating, T. J., and Zuber, A. (2012), Modelling future changes in surface ozone: a parameterized approach, Atmos. Chem. Phys., 12, 2037-2054, doi:10.5194/acp-12-2037-2012.

Wu, S., B. N. Duncan, D. J. Jacob, A. M. Fiore, and O. Wild (2009), Chemical nonlinearities in relating intercontinental ozone pollution to anthropogenic emissions, Geophys. Res. Lett., 36, L05806, doi:10.1029/2008GL036607.

Yarwood, G., Rao, S., Yocke, M., and Whitten, G. (2005), Updates to the carbon bond chemical mechanism: CB05. Final report to the US EPA, EPA Report Number: RT-0400675.

Zhang, L., Jacob, D. J., Boersma, K. F., Jaffe, D. A., Olson, J. R., Bowman, K. W., Worden, J. R., Thompson, A. M., Avery, M. A., Cohen, R. C., Dibb, J. E., Flock, F. M., Fuelberg, H. E., Huey, L. G., McMillan, W. W., Singh, H. B., and Weinheimer, A. J. (2008), Transpacific transport of ozone pollution and the effect of recent Asian emission increases on air quality in North America: an integrated analysis using satellite, aircraft, ozonesonde, and surface observations, Atmos. Chem. Phys., 8, 6117-6136, doi:10.5194/acp-8-6117-2008.

Zhang, L., Jacob, D. J., Kopacz, M., Henze, D. K., Singh, K., and Jaffe, D. A. (2009), Intercontinental source attribution of ozone pollution at western U.S. sites using an adjoint method, Geophys. Res. Lett., 36, L11810, doi: 10.1029/2009GL037950.

Zhang, L., D. J. Jacob, N. V. Downey, D. A. Wood, D. Blewitt, C. C. Carouge, A. van Donkelaar, D. B. A. Jones, L. T. Murray, and Y. Wang (2011), Improved estimate of the policy-relevant background ozone in the United States using the GEOS-Chem global model with 1/2°×2/3° horizontal resolution over North America, Atmos. Environ., 45, 6769–6776, doi: 10.1016/j.atmosenv.2011.07.054.

Zhang, Q., Yuan, B., Shao, M., Wang, X., Lu, S., Lu, K., Wang, M., Chen, L., Chang, C.-C., and Liu, S. C. (2014), Variations of ground-level O_3 and its precursors in Beijing in summertime between 2005 and 2011, Atmos. Chem. Phys., 14, 6089-6101, doi:10.5194/acp-14-6089-2014.

Zhang, Y., Y. Chen, G. Sarwar, and K. Schere (2012), Impact of gas-phase mechanisms on Weather Research Forecasting Model with Chemistry (WRF/Chem) predictions: Mechanism implementation and comparative evaluation, J. Geophys. Res., 117, D01301, doi:10.1029/2011JD015775.

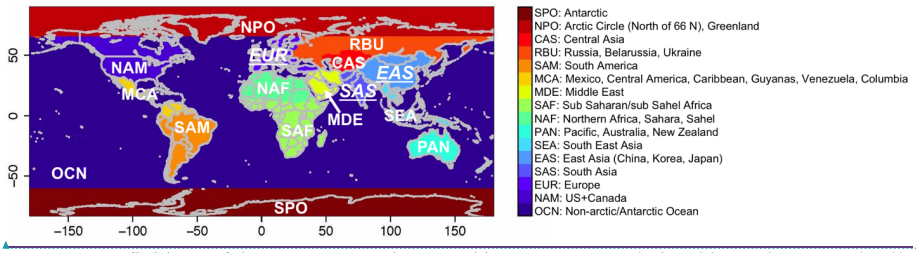


Figure 1. Definitions of the 16 source regions used in HTAP2 SR relationship study (More details in Koffi et al., 2016). The map is plotted based on data on a 0.1°×0.1° resolution grid. We focus in this study on the impact of anthropogenic pollution from selected non-North American source regions (i.e., EAS, SAS, and EUR), whose names are underlined and in italic.

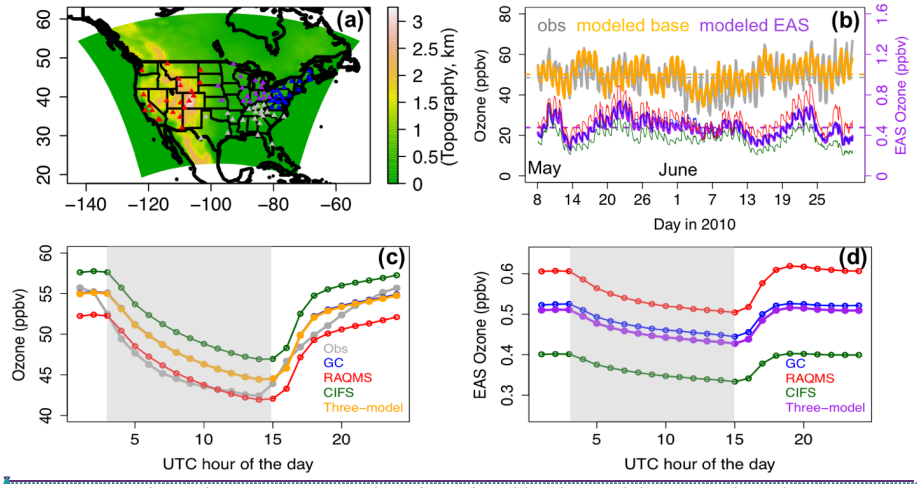
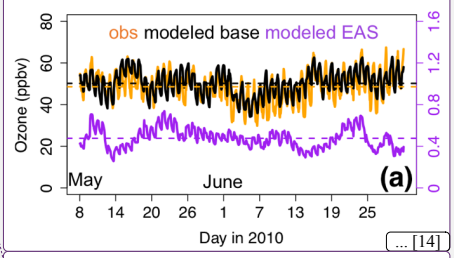


Figure 2. (a) The 60 km STEM NAM domain, colored by the model topography. The CASTNET sites used in the STEM base O_3 evaluation are marked as triangles in different colors that identify the subregions they belong to (red: western US; grey: southern US; purple: Midwest; blue: northeastern US). **(b)** Evaluation of the STEM modeled (averaged from the three base simulations using the GEOS-Chem, ECMWF C-IFS, and RAQMS base runs as the chemical boundary conditions) hourly O_3 at the western US (i.e., EPA regions 8, 9, and 10) CASTNET sites. Observations, modeled base O_3 and the modeled R(O_3, EAS, 20%) are in grey, orange, and purple lines, respectively. The horizontal dashed lines indicate the period mean values. The R(O_3, EAS, 20%) values from STEM calculations using three different chemical boundary conditions are shown separately in thin lines (blue: GEOS-Chem; red: RAQMS; green: C-IFS). The period-mean diurnal variability of the STEM modeled **(c)** base and **(d)** R(O_3, EAS, 20%) at the western US CASTNET sites. The STEM calculations using three different chemical boundary conditions are shown separately as well as averagely. Light grey-shaded areas indicate the local standard nighttime (from 6/7 pm to 7/8 am).

Formatted: Font:Bold

Deleted:



... [14]

Deleted: .

Deleted: as highlighted

Deleted:

... [15]

Moved down [16]: The horizontal dashed lines indicate the period mean values.

Formatted: Font:Bold

Deleted: (b

Moved (insertion) [16]

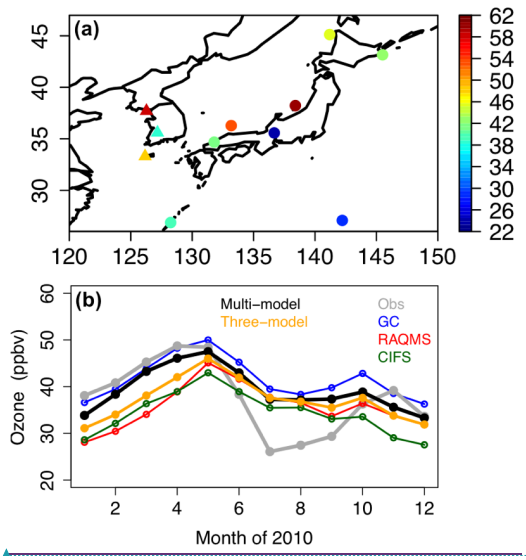


Figure 3. (a) May-June 2010 period mean surface O_3 observations in ppbv at eight Japanese (filled circles) and three Korean (filled triangles) EANET sites. **(b)** Observed and modeled monthly-mean surface O_3 in 2010 at all eleven EANET sites. The "Multi-model" and "Three-model" in the legend indicate the mean values of all eight global models and only of the three boundary condition models, respectively.

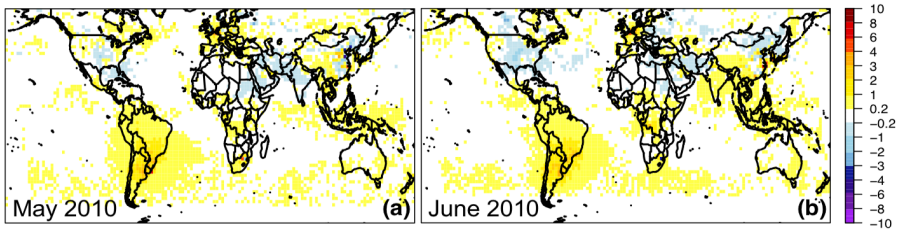


Figure 4. Evaluation of the GEOS-Chem adjoint base NO_2 product (recorded at near the satellite overpassing time) with the OMI NO_2 columns. The differences between OMI and GEOS-Chem (OMI-modeled) tropospheric NO_2 columns ($\times 10^{15}$ molec./cm^2) are shown for **(a)** May and **(b)** June 2010. Details of the comparison are included in Section 2.3.2.

Formatted: Font:Bold

Formatted: Font:Bold

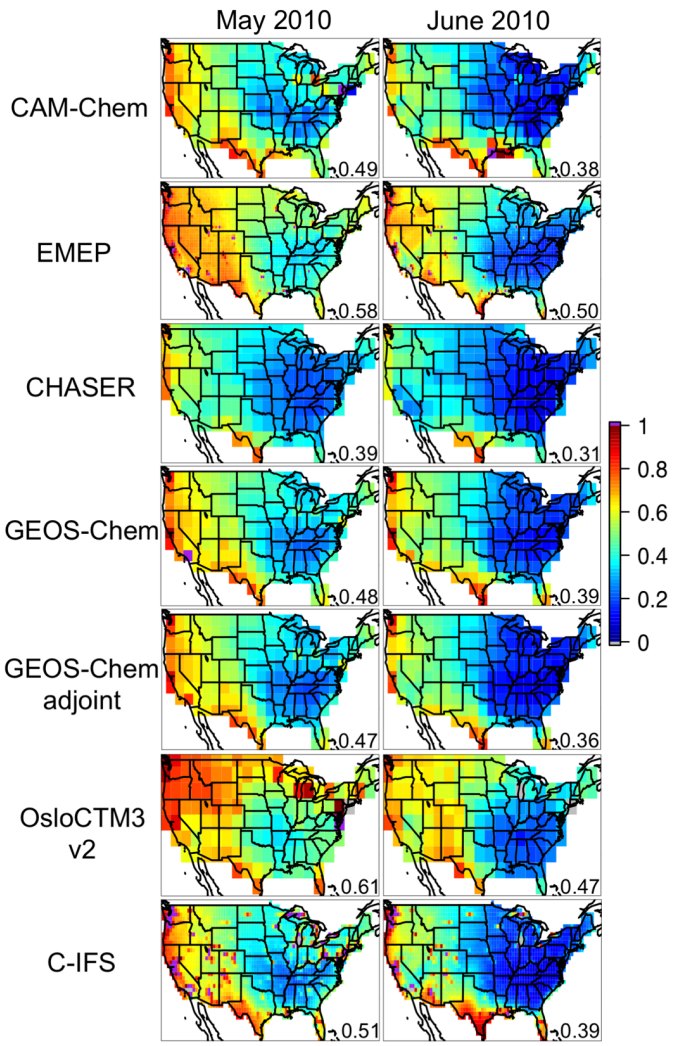
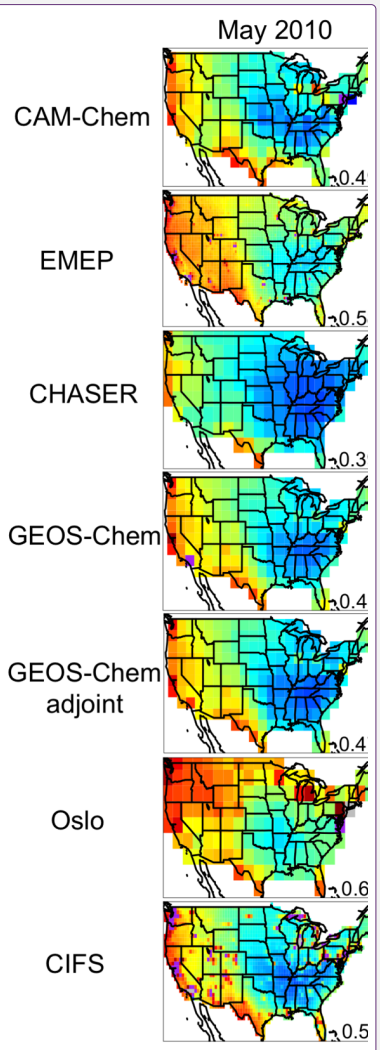


Figure 5. The RERER maps in May (left) and June (right) 2010 over the continental US, calculated based on the monthly mean O_3 from multiple global models' base and emission sensitivity simulations. The RERER metric (unitless) was defined in eq. (2) in the text. Values larger than 1 and smaller than 0 are shown in purple and grey, respectively. The US (including continental US as well as Hawaii which is not shown in the plots) mean values are indicated for each panel at the lower right corner. All models show declining RERER values from May to June, and the 7-model mean RERER values for May and June 2010 are ~0.5 and ~0.4, respectively.

Formatted: Font:Bold



Deleted:

Deleted: 4

Deleted: The

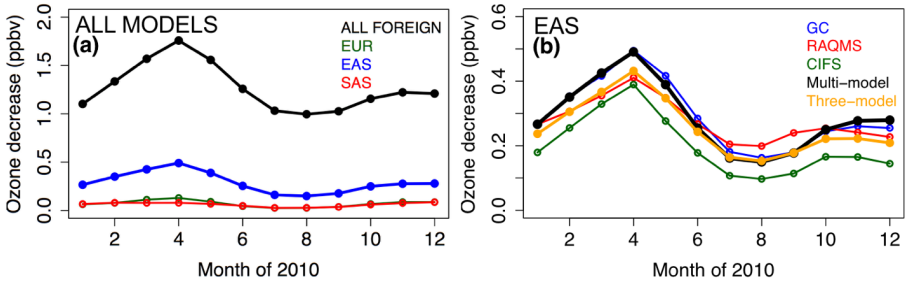


Figure 6. (a) North American (130-65°W; 20-50°N) mean O_3 sensitivity to 20% anthropogenic emission reductions in various non-North American regions, averaged from multiple (six-eight, see details in text) global models. **(b)** North American surface R(O_3, EAS, 20%) values, as estimated by single (the three STEM boundary condition models) or multi- global model means. The "Multi-model" and "Three-model" in the legend indicate the mean sensitivities of all eight global models and only of the three boundary condition models, respectively.

Deleted: 5

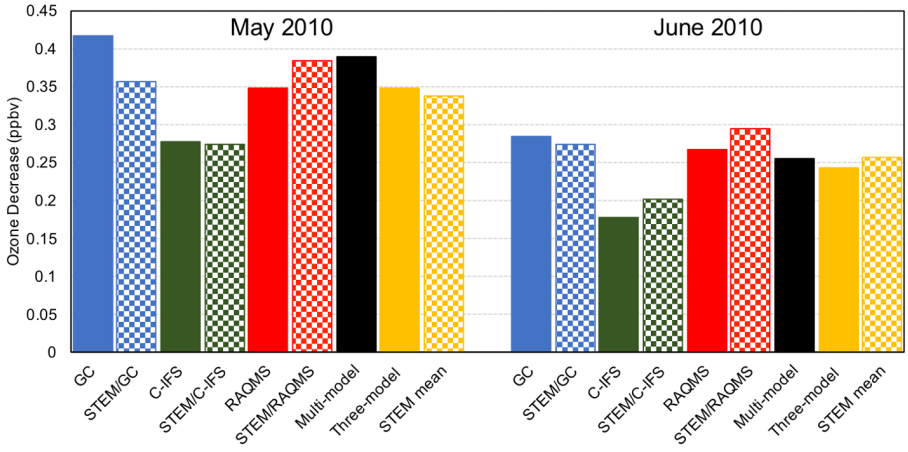


Figure 7. Monthly-mean North American (130-65°W; 20-50°N) surface R(O_3, EAS, 20%) values from multiple global and regional model simulations for May (left) and June (right) 2010. STEM model mean values were calculated from its hourly output from 8 May and on. The "Multi-model" and "Three-model" in the legend indicate the mean sensitivities of all eight global models and only of the three boundary condition models, respectively.

Deleted: 6

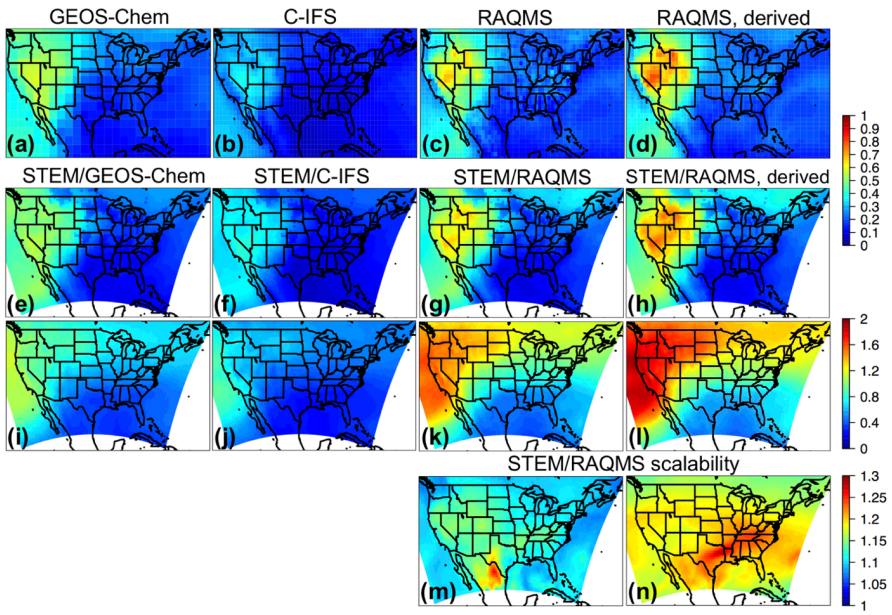


Figure 8. The monthly-mean R(O_3, EAS, 20%) in June 2010 for: **(a-d)** surface O_3 (ppbv) from the three boundary condition models, **(e-h)** STEM surface O_3 (ppbv), and **(i-l)** STEM column O_3 ($\times 10^{16}$ molecules/cm^2). R(O_3, EAS, 20%) values from the simulations associated with GEOS-Chem, ECMWF C-IFS, and RAQMS are shown in **(a;e;i)**, **(b;f;j)** and **(c;g;k)**, respectively. **(d;h;l)** show 1/5 of the R(O_3, EAS, 100%) from the simulations related to RAQMS. STEM/RAQMS-based "Scalability" S_{O3} (eq. (3)) values over the NAM are shown for **(m)** surface and **(n)** column O_3.

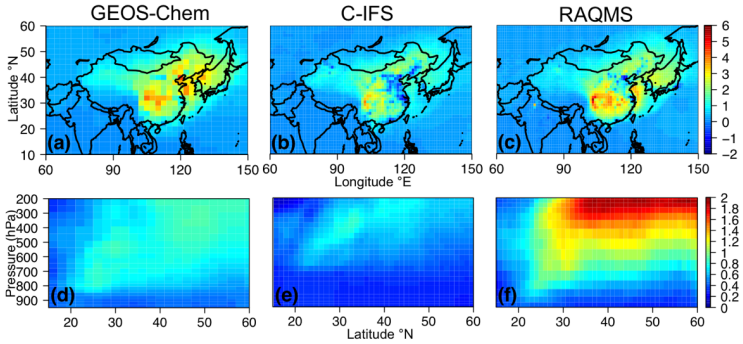
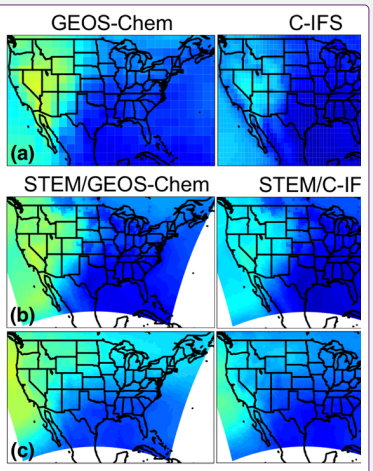


Figure 9. The monthly-mean R(O_3, EAS, 20%) in ppbv in June 2010 from the three boundary condition models at the source and near the receptor regions: **(a-c)** surface O_3 in the East Asia; and

Formatted: Font:Bold



Deleted: Figure 7.

Moved (insertion) [17]

Deleted: b

Deleted: c

Deleted: Columns 1-3 show

Deleted: ,

Deleted: Column 4 shows

Deleted: **(d)** The

Deleted: of

Deleted: (left)

Deleted: (right) in June 2010

Moved up [17]: Figure 8.

(d) O_x (GEOS-Chem) or **(e-f)** O_3 (ECMWF C-IFS and RAQMS) along the cross section of 135°W (near the west boundary of the STEM model domain as defined in Figure 2a).

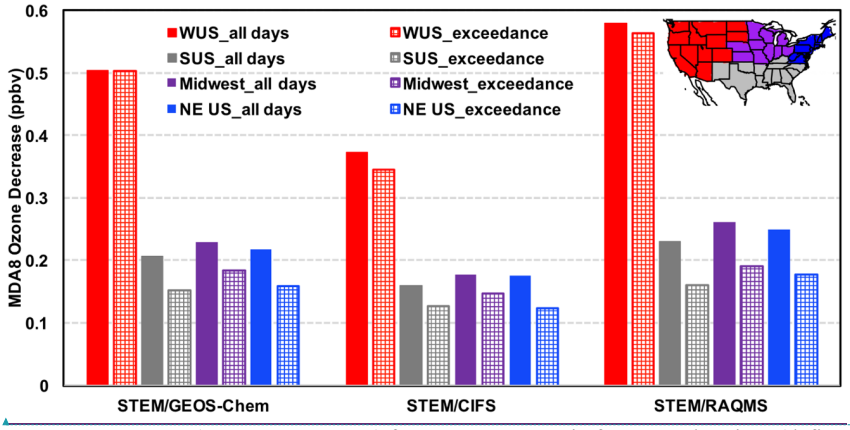


Figure 10. STEM R(MDA8, EAS, 20%) for May-June 2010 in four US subregions (defined in the inset panel, also consistent with the definitions in Figures 2/S4 and Tables 2-3), averaged on all days (bars with solid fill) and only on the days when the simulated total MDA8 O_3 concentrations were over 70 ppbv (bars with grid pattern fill). The results from the STEM runs using GEOS-Chem, ECMWF C-IFS and RAQMS boundary conditions are shown separately.

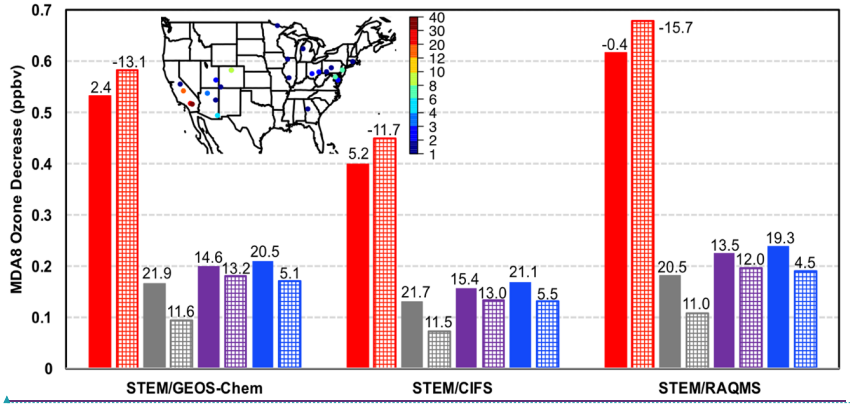


Figure 11. STEM R(MDA8, EAS, 20%) for May-June 2010 at the CASTNET sites in four US subregions (same definition as in Figure 10 inset), averaged on all days (bars with solid fill) and only on the days when the observed MDA8 O_3 concentrations were over 70 ppbv (bars with grid pattern fill). The results from the STEM runs using GEOS-Chem, ECMWF C-IFS and RAQMS boundary conditions are shown separately. Biases for the corresponding model base runs are shown above the bar plots. Inset shows at various CASTNET sites the number of days when the observed MDA8 O_3 concentrations were over 70 ppbv.

Deleted: 2b
Formatted: Font:Bold
Deleted: 9
Deleted: of
Deleted: S3
Formatted: Font:Bold

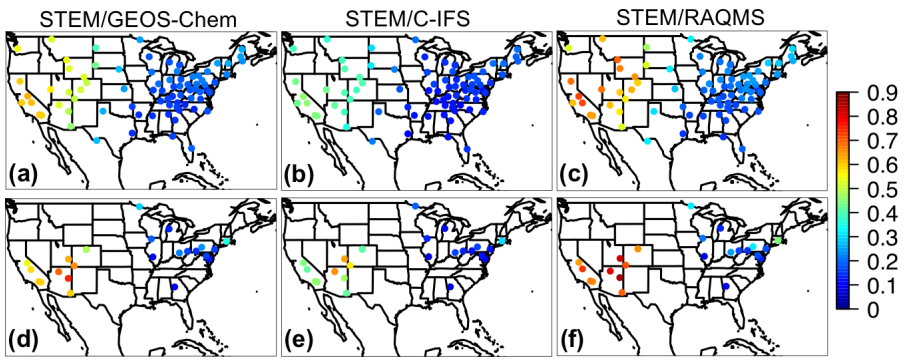


Figure 12. STEM R(MDA8, EAS, 20%) in ppbv for May-June 2010 at the CASTNET sites on **(a-c)** all days and **(d-f)** the days when the observed MDA8 O_3 concentrations were over 70 ppbv. The results from the STEM runs using **(a;d)** GEOS-Chem, **(b;e)** ECMWF C-IFS and **(c;f)** RAQMS boundary conditions are shown separately.

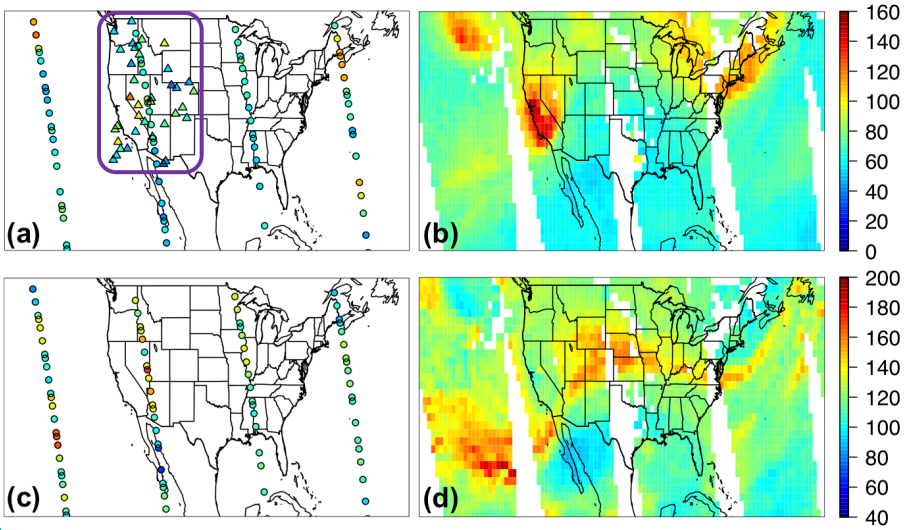
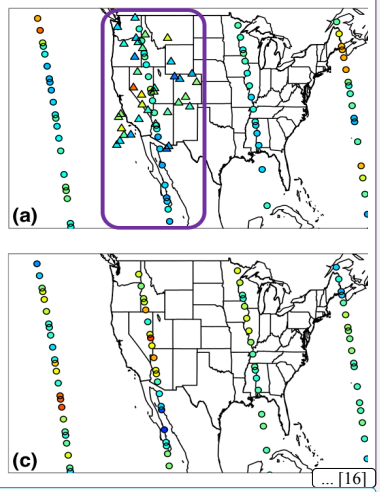


Figure 13. Case study of 9 May 2010: **(a-b)** Ozone (ppbv) and **(c-d)** CO (ppbv) at ~500 hPa from the L2 **(a;c)** TES retrievals (circles) and **(b;d)** L3 AIRS products at early afternoon local time. The L2 IASI O_3 (ppbv) at ~500 hPa retrieved using the TES algorithm (details in Section 2.3.2) at the mid- morning local times is shown on panel (b) as triangles. The O_3 profiles within the purple box in panel (a) were used in the model evaluation shown in Figure 14.

Formatted: Font:Bold



Deleted: ... [16]

Formatted: Font:Bold

Deleted: 11

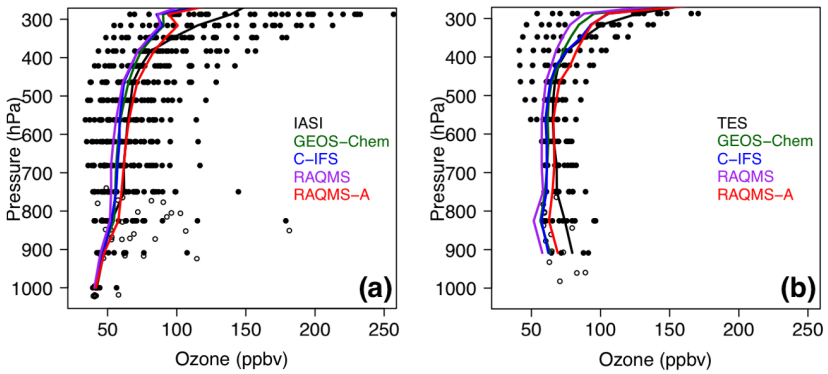


Figure 14, Case study of 9 May 2010: The comparisons between **(a)** IASI and **(b)** TES O_3 in the western US with the simulated O_3 in the STEM runs using the GEOS-Chem (green), C-IFS (blue), RAQMS (purple), and assimilated RAQMS (red) boundary conditions. The O_3 profiles within the purple box in Figure 10a were used in the evaluation. Observation operators were applied in the comparisons (details in Section 2.3.2). Solid and open dots are TES/IASI data at the TES retrieval reporting levels and at the variable surface pressure levels, respectively. Solid lines are median O_3 profiles from the satellite observations and the different STEM simulations, calculated only on the TES retrieval reporting levels.

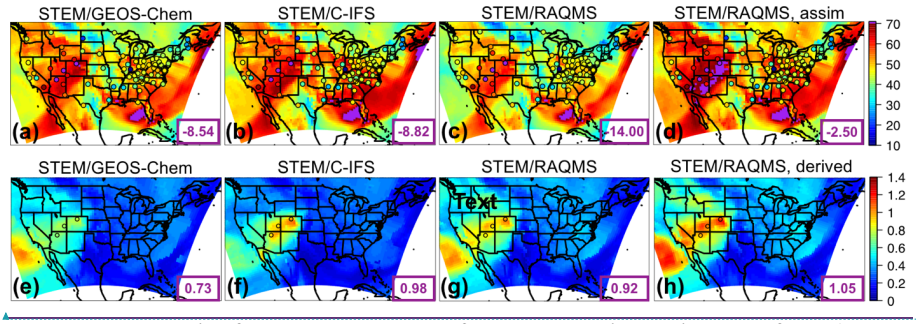
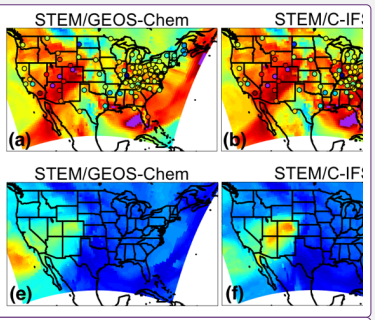


Figure 15, Case study of 9 May 2010: **(a-d)** Surface MDA8 total O_3 and **(e-h)** surface R(MDA8, EAS, 20%) from the STEM simulations using the **(a;e)** GEOS-Chem, **(b;f)** ECMWF C-IFS, and **(c;g)** RAQMS free run as the boundary conditions. **(d)** Surface MDA8 total O_3 in a STEM base simulation using the RAQMS assimilation run as the boundary conditions. CASTNET observations are overlaid in filled circles in panels (a-d). **(h)** 1/5 of the surface R(MDA8, EAS, 100%) from STEM/RAQMS simulations. The conditions at ~400-500 hPa are shown in Figure S5. Purple numbers at the lower right corners of **(a-d)** and **(e-h)** are mean model biases and mean R(MDA8, EAS, 20%) values in ppbv at the three mountain sites (Grand Canyon NP, AZ; Canyonlands NP, UT; and Rocky Mountain NP, CO) where O_3 exceedances were observed on this day. The locations of these sites are shown in panel **(e-h)** as open circles.

Deleted: 11

Formatted: Font:Bold



Deleted:

Deleted: 12

Deleted: S4

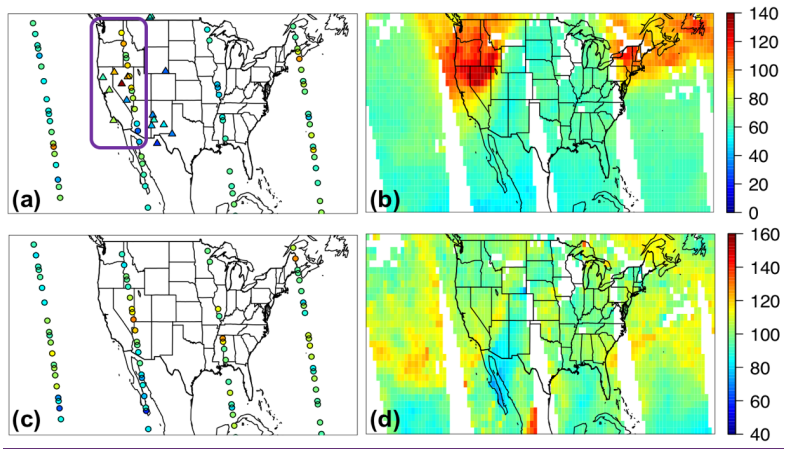


Figure 16. Same as Figure 13, but for a case study of 10 June 2010.

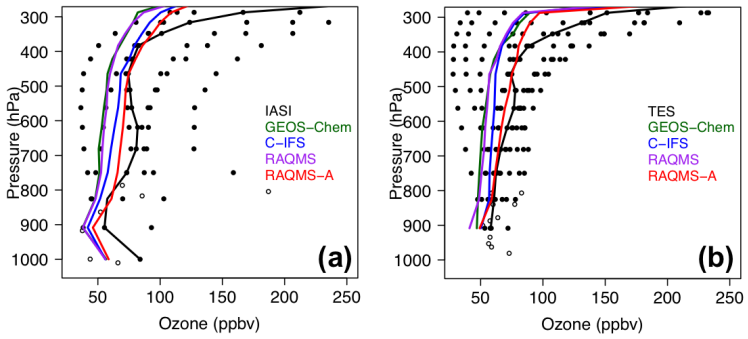
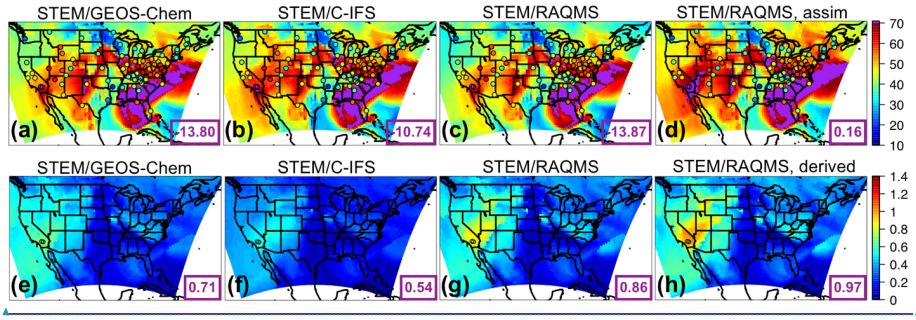


Figure 17. Same as Figure 14, but for a case study of 10 June 2010.



Formatted: Font:Bold

Figure 18. Same as Figure 15, but for a case study of 10 June 2010. The CASTNET sites with O_3 exceedances on this day are Converse Station and Joshua Tree NP in southern California.

Table 1a. HTAP2 base and sensitivity simulations by various global models. The STEM boundary condition models are highlighted in bold.

Global model, Resolution: lon×lat×vertical layer, (References)	BASE	EASALL (-20%)	EASALL (-100%)	GLOALL (-20%)	NAMALL (-20%)	EURALL (-20%)	SASALL (-20%)
CAM-Chem, 2.5°×1.9°×56 (Tilmes et al., 2016)	✓	✓		✓	✓	✓	✓
CHASER T42, ~2.8°×2.8°×32 (Sudo et al., 2002)	✓	✓		✓	✓	✓	✓
EMEP rv48, 0.5°×0.5°×20 (Simpson et al., 2012)	✓	✓		✓	✓	✓	✓
SNU GEOS-Chem v9-01-03, 2°×2.5°×47 (Park et al., 2004; http://iek8wikis.iek.fz-juelich.de/HTAPWiki/WP2.3?action=AttachFile&do=view&target=_README_GEOS-Chem.pdf)	✓	✓		✓	✓		
CU-Boulder GEOS-Chem adjoint v35f, 2°×2.5°×47 (Henze et al., 2007)	✓	✓		✓	✓	✓	✓
RAQMS, 1°×1°×35, free running (Pierce et al., 2007, 2009)	✓	✓	✓				
RAQMS, 1°×1°×35, with satellite assimilation (Pierce et al., 2007, 2009)	✓						
OsloCTM3 v2, ~2.8°×2.8°×60 (Søvde et al., 2012)	✓	✓		✓	✓	✓	✓
ECMWF C-IFS, ~0.7°×0.7°×54/1.125°×1.125°×54, as the STEM chemical boundary conditions (Flemming et al., 2015)	✓	✓		✓	✓	✓	✓

Acronyms:
CAM-Chem: Community Atmosphere Model with Chemistry
C-IFS: Composition-Integrated Forecasting System
ECMWF: European Center for Medium range Weather Forecasting
EMEP: European Monitoring and Evaluation Programme
GEOS-Chem: Goddard Earth Observing System with Chemistry
RAQMS: Realtime Air Quality Modeling System
SNU: Seoul National University

Deleted: [18]
Deleted: Relevant references for the RAQMS model [19]
Formatted [20]
Formatted Table [21]
Formatted [23]
Formatted [25]
Formatted [27]
Formatted [29]
Formatted [31]
Formatted [33]
Formatted [22]
Deleted: -
Formatted [24]
Deleted: %
Deleted: -
Formatted [26]
Deleted: %
Deleted: -
Formatted [28]
Deleted: %
Deleted: -
Formatted [30]
Deleted: %
Deleted: -
Formatted [32]
Deleted: %
Deleted: -
Formatted [34]
Deleted: %
Deleted: and horizontal resolution
Formatted [35]
Deleted: °
Formatted [36]
Deleted: °
Formatted [37]
Deleted: °
Formatted [38]
Deleted: °
Formatted [39]
Deleted: °
Formatted [40]
Deleted: °,
Formatted [41]
Formatted [42]
Deleted: °, w/
Formatted [43]
Deleted: Oslo
Formatted [44]
Deleted: °
Formatted [45]
Deleted: CIFS
Formatted [46]
Deleted: °/
Formatted [47]
Deleted: ° (used
Formatted [48]
Formatted [49]

Table 1b. STEM regional simulations for HTAP2

Boundary condition model, Resolution: lon×lat×vertical layer	BASE	EASALL (-20%)	EASALL (-100%)
SNU GEOS-Chem v9-01-03, 2°×2.5°×47	✓	✓	
RAQMS, 1°×1°×35, free running	✓	✓	✓
RAQMS, 1°×1°×35, with satellite assimilation	✓		
ECMWF C-IFS, 1.125°×1.125°×54	✓	✓	

Table 1c. STEM and its boundary condition models' key inputs and chemical mechanisms, with references. More details on the models can be found in Table 1a and the text.

Model	Meteorology	Biogenic VOCs; NO_x	Lightning	Biomass Burning	Chemical Mechanism
GEOS-Chem	GEOS-5	MEGAN v2.1 (Guenther et al., 2012); Wang et al., 2009	based on GEOS-5 deep convective cloud top heights and climatological observations (Murray et al., 2012)	GFED v3.0 (van der Werf et al., 2010)	GEOS-Chem standard NO_x-O_x-hydrocarbon-aerosol (http://acmg.seas.harvard.edu/geos/doc/archive/man.v9-01-03/appendix_1.html)
RAQMS	Online (Pierce et al., 2007)				CB-IV (Gery et al., 1989) with adjustments
ECMWF C-IFS	IFS	MEGAN-MACC, (Sindelarova et al., 2014); POET database for 2000 (Granier et al., 2005)	based on IFS convective precipitation (Meijer et al., 2001)	GFAS v1.0 (Kaiser et al., 2012)	CB05 (Yarwood et al., 2005)
STEM	WRF-ARW v3.3.1	WRF-MEGAN v2.1	based on scaled WRF convective precipitation	FINN v1.0 (Wiedinmyer et al., 2011)	SAPRC99 (Carter, 2000)

Acronyms:
CB: Carbon Bond
FINN: Fire INventory from NCAR
GFAS: Global Fire Assimilation System
GFED: Global Fire Emissions Database
IFS: Integrated Forecasting System
MACC: Monitoring Atmospheric Composition and Climate
MEGAN: Model of Emissions of Gases and Aerosols from Nature
POET: Precursors of Ozone and their Effects in the Troposphere

Formatted: Justified, Level 1, Indent: Left: 0", Suppress line numbers

Deleted:
Deleted: -
Deleted: %
Deleted: -
Deleted: %
Deleted: °
Deleted: °,
Deleted: °, w/
Deleted: CIFS
Deleted: °

WRF-ARW: Advanced Research Weather Research and Forecasting Model

Table 2a. Evaluation of the period mean (1 May-30 June, 2010) multi- global model free simulations against the CASTNET observations, only at the sites where 95% of the hourly O_3 observations are available. Evaluation of the individual models is summarized in Table 2b.

Subregion	US EPA regions contained	Number of sites	Mean bias (ppbv)		RMSE (ppbv)	
			3 BC[a] models	8 global models	3 BC models	8 global models
Western US	8, 9, 10	19	-5.68	-2.52	10.37	7.05
Southern US	4, 6	18	11.61	10.24	13.62	11.96
Midwest	5, 7	13	8.03	7.66	9.16	8.67
Northeast	1, 2, 3	17	9.55	10.63	10.28	11.24
All	1-10	67	5.49	6.22	11.11	9.96

[a]BC: Boundary Conditions

Table 2b. Evaluation of the period mean (May-June 2010) global model free simulations against the EANET and CASTNET observations. The STEM boundary condition models are highlighted in bold.

Network	Number of sites	RMSE (ppbv)							
		CAM-Chem	EMEP	CHASER	**SNU GEOS-Chem**	GEOS-Chem adjoint	**RAQMS**	OsloCTM3 v2	**C-IFS**
CASTNET	67	13.30	11.61	15.43	**15.55**	13.48	**9.32**	11.05	**11.00**
EANET	11	10.38	9.96	11.39	**9.18**	11.04	**8.60**	12.97	**10.86**

Table 2c. Evaluation of the period mean (May-June 2010) multi- global model free simulations against the EANET observations in Japan and Korea. Evaluation of the individual models is summarized in Table 2b.

Country	Number of sites	Mean bias (ppbv)		RMSE (ppbv)	
		3 BC[a] models	8 global models	3 BC models	8 global models
Japan	8	0.36	1.01	8.77	9.25
Korea	3	1.14	3.98	8.37	10.51
All	11	0.57	1.82	8.66	9.61

[a]BC: Boundary Conditions

Deleted: ... [50]

Formatted Table

Formatted: Font:Bold

Formatted: Justified

Table 3a. Evaluation of the hourly STEM simulated total O_3 (averaged from the three base simulations that used the different free-running boundary conditions) against the CASTNET surface observations for 8 May-30 June, 2010. The subregional mean R(O_3, EAS, 100%) and its correlation coefficient with the observed O_3 are also shown.

Subregion	US EPA regions contained	Number of sites	Mean elevation (km): actual/model	Mean bias (ppbv)	RMSE (ppbv)	Correlation (model base; obs)	Correlation (obs; modeled EAS)	Mean EAS sensitivity (ppbv)
Western US	8, 9, 10	22	1.75/ 1.71	1.60	4.86	0.76	0.34	0.48
Southern US	4, 6	22	0.38/ 0.31	20.33	22.13	0.58	0.27	0.15
Midwest	5, 7	16	0.29/ 0.28	15.64	17.97	0.70	0.15	0.17
Northeast	1, 2, 3	20	0.36/ 0.26	20.94	24.16	0.47	0.17	0.21
All	1-10	80	0.73/ 0.68	16.17	18.30	0.66	0.13	0.20

Table 3b. Evaluation of the hourly STEM simulated total O_3 (separately for three base simulations that used the different free-running boundary conditions) against the CASTNET surface observations for 8 May-30 June, 2010.

Subregion	US EPA regions contained	Number of sites	Mean bias (ppbv)/RMSE (ppbv)/Correlation (model base; obs)		
			SNU GEOS-Chem	C-IFS	RAQMS
Western US	8, 9, 10	22	1.68/4.83/0.77	4.16/6.63/0.70	-1.03/4.81/0.76
Southern US	4, 6	22	21.18/22.94/0.57	20.34/22.07/0.60	19.48/21.45/0.56
Midwest	5, 7	16	15.77/18.17/0.70	16.41/18.46/0.72	14.73/17.35/0.69
Northeast	1, 2, 3	20	21.25/24.36/0.47	21.86/24.80/0.48	19.71/23.40/0.45
All	1-10	80	16.57/18.62/0.66	16.89/18.84/0.67	15.03/17.52/0.64

Deleted: 3

Deleted: %), as well as

Formatted Table

Formatted: Justified

Table 4. The ranges and standard deviations (ppbv, separated by ";") of R(O_3, *source region*, 20%) by 6-8 global models (defined in eq. (1a-d)), summarized by months in 2010. The monthly multi-model mean values are shown in Figures 5-6.

Month/ Source region	All Foreign/ Non-NAM (ppbv)	EUR (ppbv)	EAS (ppbv)	SAS (ppbv)
Jan	0.38-1.69; 0.41	0.002-0.12; 0.05	0.02-0.72; 0.24	0.001-0.11; 0.04
Feb	0.92-2.07; 0.37	0.02-0.15; 0.05	0.16-0.91; 0.28	0.02-0.12; 0.04
Mar	1.30-2.37; 0.38	0.07-0.21; 0.06	0.24-1.03; 0.30	0.03-0.12; 0.03
Apr	1.42-2.46; 0.33	0.09-0.23; 0.05	0.33-1.07; 0.28	0.04-0.12; 0.03
May	1.24-1.91; 0.21	0.06-0.17; 0.04	0.24-0.75; 0.19	0.05-0.11; 0.02
Jun	1.03-1.41; 0.13	0.03-0.07; 0.02	0.14-0.39; 0.09	0.04-0.07; 0.01
Jul	0.86-1.18; 0.13	0.02-0.04; 0.01	0.08-0.22; 0.06	0.01-0.04; 0.01
Aug	0.80-1.19; 0.13	0.01-0.04; 0.01	0.07-0.20; 0.05	0.02-0.04; 0.01
Sep	0.85-1.18; 0.13	0.03-0.05; 0.01	0.10-0.25; 0.06	0.02-0.06; 0.01
Oct	0.96-1.31; 0.14	0.04-0.10; 0.02	0.17-0.42; 0.09	0.03-0.08; 0.02
Nov	0.90-1.48; 0.19	0.05-0.15; 0.04	0.17-0.54; 0.14	0.04-0.10; 0.02
Dec	0.73-1.67; 0.29	0.03-0.18; 0.05	0.14-0.66; 0.19	0.04-0.12; 0.03