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.