Response to Referee #1:

General comments

This manuscript uses the HTAP2 models to quantify source-receptor relationships for surface ozone and fine particulate matter for emission reductions occurring in six world regions and globally, as well as within three emission sectors. These source-receptor relationships are then combined with concentration-response functions to estimate premature mortality due to intercontinental, within-region, and global emissions (which includes for three separate sectors). This manuscript builds on an existing body of literature coming out of HTAP1, and so, while not particularly novel in terms of methodology, it provides an important benchmark for comparison with earlier and future work.

Thank you for your careful review of our paper and constructive comments.

A serious weakness in the paper is the absence of model comparison to observations. At the very least the paper should include a summary of any evaluation of the HTAP2 models that may be appearing in other articles in this special issue, preferably ones that are already published. A stronger paper would evaluate the specific exposure metrics used to calculate health impacts. For example, observational estimates could be added to Table 1 for regions with ground-level networks. This seems especially relevant in light of the large discrepancies across the HTAP2 models. If some models could be discarded as unrealistic, it is possible that the uncertainty in the estimated numbers of premature mortalities due to the inter-model range may decrease.

Response:

Thank you for this comment. We had previously anticipated that other HTAP2 studies would include this comparison with observations. But we now see that while two papers do include comparisons in some regions, a full global comparison with observations for all of the models used in this study is desirable here. We have now included this model evaluation with ground level observations as described in the new section 2.2 (Lines 241-294):

"Measurements from multiple observation networks are employed in this study to evaluate the model performance around the world. We evaluate model performance for the 2010 baseline simulation for 11 TF-HTAP2 models for O_3 and 8 for $PM_{2.5}$ (Table S1). For O_3, we use ground level measurements from 2010 at 4,655 sites globally, collected by the Tropospheric Ozone Assessment Report (TOAR) (Schultz et al., 2017; Young et al., 2018). The TOAR dataset identifies stations as urban, rural and unclassified sites (Schultz et al., 2017). Model performance is evaluated for the average of daily 1-h maximum O_3 concentrations for the 3

consecutive months (3m1hmaxO_3) with the highest concentrations in each grid cell, including models that only report daily or monthly O_3 as described above. This metric for O_3 differs slightly from the 6-month average of daily 1-h maximum metric used for health impact assessment, and is chosen because TOAR reports the 3-month metric but not the 6-month metric. For $PM_{2.5}$, we compare the annual average $PM_{2.5}$, using $PM_{2.5}$ observations from 2010 at 3,157 sites globally selected for analysis by the Global Burden of Disease 2013 (GBD2013) (Forouzanfar et al., 2016). Statistical parameters including the normalized mean bias (NMB), normalized mean error (NME), and correlation coefficient (R) are selected to evaluate model performance.

Table S2 and S3 present statistical parameters of model evaluation for O_3 and $PM_{2.5}$, and Figures S3-S10 show the spatial O_3 and $PM_{2.5}$ evaluation as NMB around the world, and in North America, Europe and East Asia. For 3m1hmaxO_3, the model ensemble mean shows good agreement with measurements globally with NMB of 7.3% and NME of 13.2%, but moderate correlation with R of 0.53 (Table S2). For individual models, 8 models (CAM-chem, CHASER_T42, CHASER_T106, EMEPrv48, GEOSCHEMADJOINT, GEOS-Chem, GFDL_AM3 and HadGEM2-ES) overestimate 3m1hmaxO_3 with NMB of 9.2% to 23% while 3 models (C-IFS, OsloCTM3.v2 and RAQMS) underestimate by -10.8% to -19.4% globally (Figure S3). In the 6 perturbation regions, the model ensemble mean is also in good agreement with the measurements, with -11.2% to 25.3% for NMB, 9.8% to 25.3% for NME, and -0.09 to 0.98 for R. The ranges of NMB for individual models are -18.1% to 32.3%, -24.1% to 21.3%, -24.5% to 45.0%, -26.4% to 24.5%, -30.5% to 20.3%, -35.3% to 5.4%, in NAM, EUR, SAS, EAS, MDE, and RBU, respectively (Figure S4-S6). Note that some regions (SAS, MDE, and RBU) have very few observations for model evaluation, making the comparison less robust. The underestimated O_3 in the western US and overestimated O_3 in eastern US in most models is very close to the model performance result of Huang et al. (2017) who compare 8 TF-HTAP2 models with CASTNET observations (Figure S4) , as well as earlier studies under HTAP1 (Fiore et al. 2009). Similarly, Dong et al. (2018) find that O_3 is overestimated in EUR and EAS by 6 TF-HTAP2 models, consistent with our ensemble mean result in these two regions (Figure S5-S6).

For $PM_{2.5}$, the model ensemble mean agrees well with measurements globally, with NMB of -23.1%, NME of 35.4%, and R of 0.77 (Table S3). For individual models, only 1 model (GEOSCHEMADJOINT) overpredicts $PM_{2.5}$ by 20.3%, while the other 7 models underpredict $PM_{2.5}$ by -60.9% to -7.4% around the world (Figure S7). In 6 perturbation regions, the model ensemble mean is also in good agreement with measurements, with ranges of NMB of -49.7% to 19.4%, 21.2% to 49.7% for NME, and 0.50 to 1.00 for R. The range of NMB for individual models are -46.6% to 13.9%, -76.0% to 31.9%, -35.0% to 49.7%, -50.4% to 29.5%, -52.6% to 31.5%, and -74.1% to -

19.8%, in NAM, EUR, SAS, EAS, MDE, and RBU, respectively (Figure S8-S10). Dong et al. (2018) shows that $PM_{2.5}$ is underestimated in EUR and EAS by 6 TF-HTAP2 models, consistent with our ensemble mean result in these two regions (Figure S9-S10). Note that many observations used are located in urban areas, and models with coarse resolution may not be expected to have good model performance. Also several models neglect some $PM_{2.5}$ species, which may explain the tendency of models to underestimate."

In the abstract, some context could be provided as to whether the numbers here are in line with earlier work.

Response:

For impacts of intercontinental transport, we compare results from TF-HTAP2 with the previous TF-HTAP (Anenberg et al., 2009; 2014) and comparable studies (West et al., 2009; Duncan et al., 2008), and for sectoral reductions, we compare with previous studies by Crippa et al (2017), Lelieveld et al. (2015) and Silva et al. (2016a). We have modified the abstract to compare regional results with previous studies (Lines 63-66):

"Our findings that most avoided O_3-related deaths from emission reductions in NAM and EUR occur outside of those regions contrast with those of previous studies, while estimates of $PM_{2.5}$-related deaths from NAM, EUR, SAS and EAS emission reductions agree well."

And we have also modified the abstract to compare sectoral impacts (Lines 75-81):

"In sectoral emission reductions, TRN emissions account for the greatest fraction (26-53% of global emission reduction) of O_3-related premature deaths in most regions, in agreement with previous studies, except for EAS (58%) and RBU (38%) where PIN emissions dominate. In contrast, PIN emission reductions have the greatest fraction (38-78% of global emission reduction) of $PM_{2.5}$-related deaths in most regions, except for SAS (45%) where RES emission dominates, which differs with previous studies in which RES emissions dominate global health impacts."

Specific comments

1.Lines 63-68. Does this mean outside of any of the six regions?

Response:

This metric was not sufficiently clear in the previous draft. We now present two estimates of the impact of intercontinental transport on mortality, from the source and receptor perspectives, which are added to Tables 5 and 6. The estimate of the impact of intercontinental transport on mortality from the receptor perspective uses

the RERER metric that was introduced in previous HTAP studies. We express extra-regional deaths, as presented in the abstract, as the total avoided deaths outside of each source regions from six source emission reductions. We modified how this is presented in the abstract (Lines 67-70):
"For six regional emission reductions, the total avoided extra-regional mortality is estimated as 6,000 (-3,400, 15,500) deaths/year and 25,100 (8,200, 35,800) deaths/year through changes in O_3 and $PM_{2.5}$, respectively."

We added text to clarify how the RERER metric is defined (Lines 384-396):

"We estimate the impacts of extra-regional emission reductions on mortality by using the Response to Extra-Regional Emission Reduction (RERER) metric defined by TF-HTAP (Galmarini et al., 2017):

$$RERER_i = \frac{R_{global} - R_{region,i}}{R_{global}} \quad (4)$$

where for a given region *i*, R_{global} is the change in mortality in the global 20% reduction simulation (GLO) relative to the base simulation, and $R_{region,i}$ is the change in mortality in response to the 20% emission reduction from that same region *i*. A RERER value near 1 indicates a strong relative influence of foreign emissions on mortality within a region, while a value near 0 indicates a weak foreign influence. We also estimate the total avoided extra-regional mortality from a source perspective as the sum of avoided deaths outside of each of the 6 source regions, and from a receptor perspective by summing $R_{global} - R_{region,i}$ for all 6 regions."

We modified how the results are presented on these issues (Lines 558-561):
"Overall, adding results from all 6 regional reductions, interregional transport of air pollution from extra-regional contributions is estimated to lead to more avoided deaths through changes in $PM_{2.5}$ (25,100 (8,200, 35,800) deaths/year) than in O_3 (6,000 (-3,400, 15,500) deaths/year), consistent with Anenberg et al. (2009; 2014)."

We modified this in the discussion (Lines 649-653):
"Also, total avoided deaths through interregional air pollution transport are estimated as 6,000 (-3,400, 15,500) deaths/year for O_3 and 25,100 (8,200, 35,800) deaths/year for $PM_{2.5}$ in this study, in contrast with 7,300 (3,600, 11,200) deaths/year for O_3 and 11,500 (8,800, 14,200) deaths/year for $PM_{2.5}$ in Anenberg et al. (2009; 2014)."

And we modified this in the conclusions (Lines 732-735):
"Reductions from all six regions in the transport of air pollution between regions are

estimated to lead to more avoided deaths through changes in $PM_{2.5}$ (25,100 (8,200, 35,800) deaths/year) than for O_3 (6,000 (-3,400, 15,500) deaths/year)."

2.Many prior studies are mentioned in the introduction. Are there any robust findings across this prior body of work?

Response:

We have modified the introduction to point out the robust findings by prior studies (Lines 140-143):

"These prior studies have consistently concluded that most avoided O_3-related deaths from emission reductions in NAM and EUR occur outside of those regions, while most avoided $PM_{2.5}$-related deaths occur within the regions."

3.Lines 246-248. Is the actual value of β given somewhere?

Response:

We have added text to give the value of RR from the Jerrett study, from which Beta is calculated from equation 1 (Lines 316-318):

"For O_3, RR = 1.040 (95% Confidence Interval, CI: 1.013-1.067) for a 10 ppb increase in O_3 concentrations (Jerrett et al., 2009), which from eq. 1 gives values for β of 0.00392 (0.00129-0.00649)."

4.Line 261. Make sure all terms in equation 3 are defined.

Response:

Burnett et al. (2014) defines the function given and specifies three parameters (α, γ, δ) which they use to allow more flexibility in fitting the cause-specific RR. These parameters therefore do not have specific physical meaning, and are used in the functional fitting, so we refer the reader to Burnett's paper to understand these parameters more fully (Lines 329-334):

"RR is calculated as:

For $z<z_{cf}$, $RR_{IER}(z) = 1$ (2)

For $z \geqq z_{cf}$, $RR_{IER}(z) = 1 + \alpha\{1 - exp[-\gamma(z - z_{cf})^{\delta}]\}$ (3)

where z is the $PM_{2.5}$ concentration in μg/m^3 and z_{cf} is the counterfactual concentration below which no additional risk is assumed, and the parameters α, γ, and δ are used to fit the function for cause-specific RR (Burnett et al., 2014)."

5.Line 267-268. Elaborate on Zcf: does it vary from 5.8 to 8.8 g/m^3 in space and time?

Response:

We have revised to clarify (Lines 338-341):

"A uniform distribution from 5.8 μg/m^3 to 8.8 μg/m^3 is used for z_{cf} as suggested by

Burnett et al. (2014), which does not vary in space nor time. For uncertainty analysis, we use results from 1,000 Monte Carlo simulations of Burnett et al. (2014) to calculate RR in each grid cell by eq.2 or eq. 3."

6.Figures S8 and S9 are referred to several times in the text but are impossible to read. I suggest splitting them each into 4 figures, with half the models on each, one for the regional perturbations and one for the sectoral perturbations. The full range of the colorbar isn't used, so consider using a different color bar that allows for one to read the values off the figure more easily.

Response:

We split Figures S8-S9 into 6 pages, and we use a different color bar to show full range of data. See the updated Figs. S14-S17.

7.Lines 318-320. Is this intended to be a quantitative comparison? If so, are the metrics reported here and in the Lin et al. studies the same?

Response:

No, we don't intend to show a quantitative comparison with Lin et al. (2012 and 2017) due to the different ozone metrics evaluated. Instead, we suggest that the ozone responses in the western US to emission reductions from EAS are similar to those of Lin et al. (2012 and 2017) who show that a model can capture the measured western US ozone increases due to rising Asian emissions. We add this text (Lines 423-426):

"Our ensemble shows similar ozone responses in the western US to emission reductions from EAS (Figs. 1c) as those modeled by Lin et al. (2012 and 2017), who show that a model can capture the measured western US ozone increases due to rising Asian emissions."

8.Lines 449-459. This seems like methodology and could be included in the earlier section.

Response:

We have moved these lines into the method section (Lines 376-382):

"We also quantify the uncertainties in mortality due to the spread of air pollutant concentrations across models, RRs, and baseline mortality rates, as contributors to the overall uncertainty, expressed as a coefficient, of variation and compare the result with the Monte-Carlo analysis estimate. To do so, we hold two variables at their mean values and change the variable of interest within its uncertainty range; for example, using mean RRs and baseline mortality rates, we analyze the spread of the model ensemble to calculate the coefficient of variation caused by model

uncertainty."

9.Lines 545-547. Could the use of a different year make a difference here?

Response:

We agree with the reviewer that the different year could be responsible for part of the differences between studies. We have revised the text (Lines 653-657):

"These differences likely result from different concentration-response functions and the use of 6 regions here vs. 4 by Anenberg et al. (2009; 2014). In addition, updated atmospheric models and emissions inputs, as well as different atmospheric dynamics in the single years chosen in TF-HTAP1 vs. TF-HTAP2 may contribute to the differences."

10.Lines 559-560. This seems like an important point and suggest including in abstract and conclusions.

Response:

We have revised the abstract to add this comment (Lines 72-75):

"For NAM and EUR, our estimates of avoided mortality from regional and extra-regional emission reductions are comparable to those estimated by regional models for these same experiments."

And we have added it to the conclusions (Lines 735-738):

"For NAM and EUR, our estimates of avoided mortality from regional and extra-regional emission reductions are comparable to those estimated by regional models in AQMEII3 (Im et al., 2018) for these same emission reduction experiments."

11.Lines 609-610. Given the large ranges, is it really meaningful to report averages?

Response:

The overall percentage is derived from all 6 regional emission reductions altogether, not the average of percentages for each region. We've revised to clarify (Lines 722-727):

"For regional scenarios, 6 source emission reductions altogether can cause 84% of the global avoided O_3-related premature deaths within the source region, ranging from 21 to 95% among 6 regions, and 16% (5 to 79%) outside of the source region. For $PM_{2.5}$, 89% of global avoided $PM_{2.5}$-related premature deaths are within the source region, ranging from 32 to 94% among 6 regions, and 11% (6 to 68%) outside of the source region."

12.Table 4. What is an "empirical mean"?

Response:

Since we conduct 1,000 Monte Carlo simulations to propagate uncertainty from baseline mortality rates, modeled air pollutant concentrations, and the RRs in the health impact functions, the mean of the result is called the "empirical mean" as the mean of 1,000 simulations. We added this explanation into Table 4:

"Empirical mean is the mean of 1,000 Monte Carlo simulations."

We also revised the method section to explain where this result is used (Lines 373-375):

"The mean of the 1,000 Monte Carlo simulations (the "empirical mean") may differ from the mean when using the mean RR."

13.Table S1. Why not calculate $PM_{2.5}$ consistently across models from the individual components?

Response:

Different models use different functions to represent $PM_{2.5}$, that are appropriate for each model based on how different species are defined in the models. We choose to use the reported $PM_{2.5}$ from each model, rather than to recalculate $PM_{2.5}$ based on their reported species concentrations. We include the functions used by each model in Table S1 to communicate the species that each model simulated, and other modeling differences, where for example some models may be missing important species, but we do not apply these functions ourselves in this study.