Response to Referee #2

General comments

This manuscript uses the multi-model results from HTAP2 project to estimate mortality for the baseline year 2010, and health benefits from reduced emissions in source regions. In general, it is well organized and written, and the multi-model results can provide more reasonable range than single model results in previous studies. However, some details are not well documented and explanations are too general, but important for readers.

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

Specific comments

1.Page 3, line 100: It is better to provide some brief explanation of reasons for large differneces in estimates (4.2 and 2.1 million premature deaths).

Response:

We have added short discussion on this point (Lines 107-109):

"These differences in GBD estimates result mainly from differences in concentration response functions and estimates of pollutant concentrations."

2.Page 5, line 159: Please specify if the perturbation is increasing or decreasing.

Response:

We reduced the anthropogenic emissions by 20% (Line 170):

"Anthropogenic emissions were reduced by 20% in six source regions: …"

3.Page 6, line 190-203: how do these models perform in simulating ozone and $PM_{2.5}$

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."

4.Page 7, line 246-257: what beta value is used in this study? any source for the used RR=1.040? Please clarify..

Response:

We added this (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)."

5.Page 8, line 264-271: The used RR framework here is not actually the latest. Please refer to Cohen et al. (2017).

Response:

Our work was nearly completed before Cohen et al. (2017) was published, and so we chose the most recent available RR from Burnett et al. (2014) for $PM_{2.5}$. This function was widely used in many studies, including by Silva et al (2016a), Lelieveld et al (2015), and GBD 2010 (Lim et al., 2012). However, we have added short discussion on this difference (Lines 473-481):

"Cohen et al. (2017) use RRs for particulate matter for IHD and stroke mortality that are modified from those used by Burnett et al (2014) and applied age modification to the RRs, fitting the IER model for each age group separately. The updated IER with estimated higher relative risks, together with greater global pollution and baseline mortality rates in the low-income and middle-income countries in east and south Asia leads to the higher absolute numbers of attributable deaths and disability-adjusted life-years in GBD 2015 than estimated in GBD 2013 (Forouzanfar et al., 2016). Also, GBD 2015 includes child lower respiratory infections estimate whereas we do not".

6.Page 8, line 276-277: Please clarify how you treat age distribution in the 2011

populaiton dataset.

Response:

We add text (Lines 355-358):

"For the population of adults aged 25 and older, we use ArcGIS 10.2 geoprocessing tools to estimate the population per 5-year age group in each cell by multiplying the country level percentage in each age group by the population in each cell."

7.Page 8, is sex difference considered in the estimation?

Response:

No, we only consider age-specific RR, as given by the health impact functions we use and the underlying epidemiological studies.

8.Page 8, line 282: Monte Carlo simulation is powerful to address uncertainty issues. However, the way of including model air pollutant concentrations is a bit misleading. The procedure in this study is actually the range of multi-model results. However, it is possible that this range deviate from the observations. Without showing model evaluation, we don't have confidence how reliable is the range from multi-models.

Response:

We've added the model evaluation in section 2.2 (Lines 241-294).

We also added an acknowledgement that the range of models in an ensemble is not a true reflection of the uncertainty in emissions to the method section (Lines 371-373):

"One should acknowledge that the range of models in an ensemble is not a true reflection of the uncertainty in emissions to concentration relationships."

10.Page 9, line 306: The texts refer to supplemental plots many times. I would suggest move some important figures from supplemental materials.

Response:

We've moved figures S6-S7 to figures 1-2 in the main paper. The order of figures has been updated to reflect this change in main paper as well as the supporting document.

11.Page 10, line 368-369: Please provide more details here: the updated baseline mortality rate in 2017, and how population is different. This comparison is too general here. In my understanding, the biggest change from GBD framework from old to latest (Cohen et al., 2017) is not just baseline mortality. In Cohen et al. (2017), the RR for stoke is totally different from previous version GBD, and LRI disease is added in addition to IHD, LC, COPD and stroke.

Response:

As stated before, we now provide details on how RRs were updated for use by Cohen et al. (2017) (Lines 473-481):

"Cohen et al. (2017) use RRs for particulate matter for IHD and stroke mortality that are modified from those used by Burnett et al (2014) and applied this age modification to the RRs, fitting the IER model for each age group separately. The updated IER with estimated higher relative risks, together with greater global pollution and baseline mortality rates in the low-income and middle-income countries in east and south Asia leads to the higher absolute numbers of attributable deaths and disability-adjusted life-years in GBD 2015 than estimated in GBD 2013 (Forouzanfar et al., 2016). Also, GBD 2015 includes child lower respiratory infections estimate whereas we do not."

12.Page 11 line 382-383: Please clarify how the avoid deaths is calculated. the IER model is not linear: at the high end large changes in pollutant will not result in large changes in death, some studies used average changes, some used marginal. How is this addressed here?

Response:

The percentage of the global change in O_3-related deaths within the source region is computed by the number of avoided deaths within source region divided by the number of avoided deaths globally from 20% source emission reduction. We've revised to clarify this calculation (Lines 495-496):

"The number of avoided deaths within source region is divided by the number of avoided deaths globally"

We added text to discuss the issue about IER model (Lines 343-352):

"However, in the IER model, the concentration–response function flattens off at higher $PM_{2.5}$ concentrations, yielding different estimates of avoided premature mortality for identical changes in air pollutant concentrations from less-polluted vs. highly-polluted regions. That is, one unit reduction of air pollution may have a stronger effect on avoided mortality in regions where pollution levels are lower (e.g., Europe, North America) compared with highly polluted regions (e.g., East Asia, India), which would not be the case for a log-linear function (Jerrett et al., 2009; Krewski et al., 2009). Therefore, using the IER model in this study may result in smaller changes in avoided mortality in highly polluted areas than using the linear model."

13.Page 11 line 406-408: The explanation here is not convincing.

Response:

We've revised this explanation to (Lines 520-522):
"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."

14.It would be great to make a table to inter-compare the response of sector reductions, which is highly uncertain from different models, and please discuss it too.
Response:
We've listed this inter-comparison between models for sector reductions in TableS9-S10 and discussed these differences in Lines 616-625:
"Considering results from individual models, we found that O_3-and $PM_{2.5}$-related mortality from TRN emission reductions show greater relative uncertainty than from PIN or RES (Table 5-6 and Table S9-S10), reflecting a greater spread of results across models. Regional impacts from individual models also differ from the ensemble mean result - e.g., for O_3, GEOSCHEMADJOINT and OsloCTM3.v2 show that reducing PIN emissions causes the greatest fraction of avoided O_3-related deaths in EUR, while GEOSCHEMADJOINT, HadGM2-ES and OsloCTM3.v2 show that TRN emissions have the greatest fraction of avoided O_3-related deaths in RBU (Figs. S20). For $PM_{2.5}$, CHASER_t42 and GEOSCHEMADJOINT show that reducing PIN emissions causes the greatest fraction of avoided $PM_{2.5}$-related deaths in SAS (Figs. S21)"

In addition, we also compare our O_3 and $PM_{2.5}$-related premature deaths attributable to PIN, TRN and RES emissions with previous studies conducted by Silva et al. (2016) and Lelieveld et al. (2015) in table 7 and discuss the differences from our estimates in Lines 601-615 :
"In comparison with other studies (Table 7), our conclusion that PIN emissions cause the most O_3-related deaths and TRN emissions cause the greatest fraction of avoided deaths in most regions agrees well with Silva et al (2016a). For $PM_{2.5}$, reducing PIN emissions avoids the most $PM_{2.5}$-related premature deaths globally (128,000 (41,600, 179,000) deaths/year) and in most regions (38-78% of the global emission reduction), except for SAS (45%) where the RES emission dominates. Although these findings differ from those of Lelieveld et al (2015) and Silva et al (2016), who find that Residential emissions have the greatest of impact on $PM_{2.5}$ mortality globally and in most regions, all studies agree that PIN emissions have the greatest impact in NAM. Our result is also comparable with Crippa et al (2017) who find that PIN emissions have the greatest health impact in most countries. Although comparable emission inventories are used (i.e. Lelieveld et al (2015) use EDGAR emissions while Silva et al (2016) use RCP8.5. emissions), our lower mortality estimate for RES emissions may

be explained by our 20% reductions relative to the zero-out method, and the different years simulated."

and Lines 674-686 :
"Differences in our estimates of premature mortality attributable to air pollution from three emission sectors (multiplied by 5) may be explained by methodological differences relative to previous studies (Silva et al., 2016; Lelieveld et al., 2015), including our use of 20% emission reductions versus the zero-out method in those studies, different emission inventories, a multi-model ensemble versus single models, and differences in baseline mortality rates, population, and concentration response functions. Our finding that TRN emissions contribute the most avoided deaths for O_3 in most regions agrees well with the result by Silva et al (2016a), but differs for $PM_{2.5}$ mortality for which we find that PIN emissions cause the most deaths, while both Silva et al (2016a) and Lelieveld et al (2015) find that RES emissions are responsible for the most deaths. This discrepancy may be explained by different $PM_{2.5}$ species included in individual models, as we showed that changes in $PM_{2.5}$ concentration to TRN emission differ across models."