We thank the reviewer for the constructive comments. We have tried to address all the points raised in the review.

Comment: Lines 84-96 should be updated with the most recent GBD 2016 numbers

Response: The numbers are updated (Lines 87-89).

Comment: Lines 118-153 could use some organization. This section is basically just listing results from individual studies without synthesizing them or connecting them to the present study. It's not clear as written by this section is there.

Response: We have now extended this section (Lines 120-132).

Comment: Line 188-190 states that this is the first study to use a common approach for health impact assessment across US and Europe, but the HTAP ozone and PM2.5 health impact assessments referenced earlier used a similar approach. Perhaps the authors are referring only to the economic valuation portion? If so, I'm still not sure this is the first study to do that since there are now several (perhaps many) global health impact and valuation studies that use a common approach for all countries/regions, including US and Europe.

Response: The economic valuation was not included in the GBD assessment and others. OECD has published a global assessment with economic valuation, but without a consistent atmospheric modelling framework.

Comment: Lines 296-298: given that this paper's focus is on the health impacts, and not the modeling, there should be much more detail given here about the health impact methods in addition to, or instead of, the modeling detail, which can be found in other places and referenced. The health methods quickly summarized here diverge from the methods used by the Global Burden of Disease, U.S. EPA, and many recently published papers. So this needs to be explained, expanded, and justified quite a bit more. As stated, summing ozone deaths with PM2.5 YOLL doesn't make logical sense, as one is cases and one is years, and what is being divided by 10.6 and why? The CAFÉ reference is 12 years old, and air pollution epidemiology and health impact assessment has advanced quite a bit since then. For ozone, there are now studies showing effects of long-term exposure on mortality, just like for PM, so why are only short-term ozone impacts calculated?

Response: EVA methodology is now extended (Lines 326-420). The selected health end-points are fairly conventional and aligned to the impact assessments that have been done for the European Commission and the European Environment Agency (EEA) up to 2013; they have been richly documented elsewhere. It was not the purpose here to develop a novel health impact assessment, or to compare in detail with GBD or US-EPA, but rather to explore its implications across the two continents.

Comment: Lines 299-302: The ERFs listed in Table 2 are quite a bit out of date, particularly for the U.S. studies. Most of these are 20 years old. There have been many studies now reporting updated ozone and PM2.5 risk estimates for the American Cancer Society cohort which can be used. And these are not necessarily consistent magnitudes compared with the old studies.

Response: These ERFs are consistent with the functions used by the EEA and conservative as they are updated only if recommended by the WHO even though there are recent studies providing updated functions. This is now added to the manuscript. A new version of the model is currently under development with more updated ERFs, additional species such as NO_2, chronic O_3-related mortality, and a breakdown of the aerosol components.

Comment: Table 2 needs concentration metrics to which each ERF applies. Section 2.2 should state which concentration metrics were drawn from the models (annual average, annual average of 8-hr daily max, etc.) used which each ERF. I see now these are indicated starting in line 376, but not explained, and should be in section 2.2.

Response: Table 2 includes which pollutants are used for each health impact. The section is also extended now to include more specifically what metric are used on what temporal resolution (Lines 358-360), following: "EVA calculates and uses the annual mean concentrations of CO, SO_2 and $PM_{2.5}$, while for O_3, it uses the SOMO35 metric that is defined as the yearly sum of the daily maximum of 8-hour running average over 35 ppb, following WHO (2013) and EEA (2017)."

Comment: Section 2.2 should also give some equations used to calculate health impacts. It's difficult to understand what was done and impossible to judge whether it's technically sound.

Response: We have now extended this section and it is now clearer on the implementation of the model (Lines 326-420).

Comment: Section 2.2 were the exposure response functions applied in a linear equation or some other functional form (e.g. log-linear)? This is important for the perturbation simulations because you are reducing pollution at the high end, where the shape of the curve can have a big impact on the magnitude of health benefits estimated.

Response: We have now added the following sentence (Lines 353-355): "EVA uses ERFs that are modelled as a linear function, which is a reasonable approximation as showed in several studies (e.g. Pope et al., 2000; the joint World Health Organization/UNECE Task Force on Health (EU, 2004; Watkiss et al., 2005))."

Comment: Section 2.2 should also indicate the source of baseline disease rates to calculate health impacts.

Response: the EVA model applies universal baseline rates from Statistics Denmark, therefore not country-specific, which is a simplification, although aligned to the Eurozone countries.

Comment: Section 2.2 did you first estimate health impacts from each individual model and then average, or first average the concentrations across models and then estimate health impacts?

Response: We have now extended the section (Lines 288-294). All modeling groups interpolate their model outputs on a common 0.25°×0.25° resolution AQMEII grid predefined for Europe (30°W - 60°E, 25°N - 70°N) and North America (130°W - 59.5°W, 23.5°N - 58.5°N). All the analyses performed in the present study use the pollutant concentrations on these final grids. Health impacts are first calculated for each individual model and then the ensemble mean, median and standard deviation are calculated for each health impact. In order to be able to estimate an uncertainty in the health impacts calculations, none of the models were removed from the ensemble.

Comment: Section 2.2 what spatial resolution was used to estimate health impacts? Part of the problem with previous studies of PM long-range transport is that the grid resolution was too coarse to adequately capture health benefits from reducing local PM. Spatial scale is important.

Response: We have now extended the section (Lines 288-294). All modeling groups interpolate their model outputs on a common 0.25°×0.25° resolution AQMEII grid predefined for Europe (30°W - 60°E, 25°N - 70°N) and North America (130°W - 59.5°W, 23.5°N - 58.5°N). All the analyses performed in the present study use the pollutant concentrations on these final grids. Health impacts are first calculated for each individual model and then the ensemble mean, median and standard deviation are calculated for each health impact. In order to be able to estimate an uncertainty in the health impacts calculations, none of the models were removed from the ensemble.

Comment: Section 3.2 are the plus/minus numbers given with all the results the range of health impacts calculated with individual models? How was uncertainty in the exposure response function accounted for?

Response: We have now added the following (Lines 291-294). Health impacts are first calculated for each individual model and then the ensemble mean, median and standard deviation are calculated for each health impact.

Comment: Line 413 appears to be missing a 0 in the HTAP2 result

Response: We have now corrected this.

Comment: Line 421 what is meant by "by construction"?

Response: We have removed this phrase.

Comment: There are many references to the Liang (in preparation) study, but since this study is not yet available the usefulness of these comparisons is limited. It is often used as justification that the present study was done right, since the numbers match up. But there is not currently enough information from either study to judge that.

Response: We have added more comparisons with other published studies (Lines 563-566; 617-620).

Comment: There are many tables with numbers for health impacts that are difficult to digest. Suggest replacing some of these with figures to highlight the most salient points.

Response: We have moved some of the tables (Table 3 and Table 4) in the supplement and kept the ensemble mean results together with the optimal ensemble results from old Table 7 to the new Table 3. However, we believe that these numbers should be explicitly presented in the manuscript as particularly the morbidity calculations are for the first time calculated for both continents and transferring them into figures would lose the details.