We thank the reviewer for the constructive comments. We have responded to all the comments in the new version of the manuscript.

Summary comments

Comment: This manuscript is an ambitious effort to simulate air quality changes and estimate health impacts using an ensemble of models. The results clearly reflect a substantial effort on the part of the authors. I have three primary concerns:

(1) the health impact assessment is insufficiently documented. In particular, the manuscript does not clearly describe the procedure for selecting and applying health endpoints to quantify or the source of the baseline incidence rates in the U.S. and Europe.

Response: 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, but rather to explore its implications across the two continents. A new generation of health impact assessments are expected to make reference to the meanwhile established WHO HRAPIE consensus guidelines.

(2) Reasonable people can disagree as to whether it's appropriate to quantify the economic value of years of life lost. However, the manuscript does not attempt to provide a rationale for this choice.

Response: This is a fairly crucial aspect of mortality impacts, which EU and USA simply approaches differently – we here adhere to the European approach, the main advocate of which was Ari Rabl (Rabl, Spadaro and Holland, 2014). See further below.

(3) Finally, the authors should indicate whether each of the air quality and health impact models used have been peer reviewed and whether the source code is publicly available.

Response: As seen in Table 1 and now in the supplementary material, there a number of CTMs used in the AQMEII exercise. Some of these CTMs are community models, such as WRF/Chem, CMAQ and CAMx, while others are not community models and being used by the main developers so that the model is not publicly available but can be shared upon collaboration. Only one health impact model has been used, using different concentration inputs from each of the CTMs. EVA system is not a community model either and developed internally by Aarhus University, but has been used upon collaboration with other institutes.

Detailed comments

Comment: Line 46: Is this correct? The outdoor air pollution portion of the Global Burden of Disease studies have applied a consistent modelling framework to predict ambient concentrations of common air pollutants, and quantify the number of premature deaths attributable to outdoor fine particles and ground-level ozone. Other studies, including Anenberg et al. (2010, 2014) quantify global ozone and PM-attributable deaths due to anthropogenic emissions.

Response: GBD does not provide economic estimates. Same for Anenberg et al. (2010 and 2014).

Comment: Line 50: Anthropgenic and non-anthropogenic?

Response: The perturbation simulations target anthrpogenic emissions. This is now added to the text (Line 49).

Comment: Line 53: Did you estimate impacts down to some background concentration, or to zero?

Response: EVA system uses a cut off value of 35 ppb to calculate health impacts from ozone and used to calculate the SOMO35 metric. Regarding PM2.5, no threshold is being applied, following the EEA recommendations (See Line 388-396).

Comment: Lines 52-65: Here and elsewhere it would be helpful to distinguish between the air quality modeling portion of the ensemble and the health impact estimation portion of the ensemble.

Response: The health impacts are calculated from each CTM individually. Therefore, the health impact ensemble includes health impacts using concentrations from the different CTMs. We have now made this more clear in the text as follows (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."

Comment: Lines 66-77: Are these a sum of the PM2.5 and ozone-related premature deaths?

Response: The numbers reflect the total premature death. The text now reads (Lines 63-71): "A total of 54 000 and 27 500 premature deaths can be avoided by a 20% reduction of global anthropogenic emissions in Europe and the U.S., respectively. A 20% reduction of North American anthropogenic emissions avoids a total premature death of ~1 000 in Europe and 25 000 total premature deaths in the U.S. A 20% decrease of anthropogenic emissions within the European source region avoids a total premature death of 47 000 in Europe. Reducing the East Asian anthropogenic emissions by 20% avoids ~2000 total premature deaths in the U.S. These results show that the domestic anthropogenic emissions make the largest impacts on premature death on a continental scale, while foreign sources make a minor contributing to adverse impacts of air pollution."

Comment: Line 85: What does "scale dependent challenge" mean in this context?

Response: We have modified the sentence to be more clear (Line 79-81): "Air pollution is a transboundary phenomenon with global, regional, national and local sources, leading to large differences in the geographical distribution of human exposure."

Comment: Line 93: Suggest updating with most current GBD published value. Lines 104-109: These two statements are difficult to reconcile.

Response: This part has been modified with newer numbers and for better readability (Lines 87-89): "The Global Burden of Disease Study 2015 estimated 254 000 O_3-related and 4.2 million anthropogenic $PM_{2.5}$-related premature deaths per year (Cohen et al., 2017)."

Comment: Line 150: This isn't quite right. That paper estimated a total of between 130k and 350k PM & O3 related deaths. Note also that this paper quantified impacts from anthropogenic emissions alone.

Response: We have now corrected the sentence as (Line 153-155): "Fann et al. (2012) calculated 130,000 - 350,000 premature deaths associated with O_3 and $PM_{2.5}$from the anthropogenic sources in the U.S. for the year 2005."

Comment: Line 155: Suggest rephrasing for clarity.

Response: We have changed the sentence as (Line XXX): "Observations have spatial limitations particularly when assessments are needed for large regions."

Comment: Lines 197-202: I had a hard time following these statements. In particular, I could not understand what exactly you did to minimize error and what redundancy you're referring to.

Response: We have now rephrased this part as follows (Lines 202-205): "Finally, following the conclusions of Solazzo and Galmarini (2015), the health impacts have been calculated using an optimal ensemble of models, determined by error minimization. This approach can assess the health impacts with reduced model bias, which we can then compare with the classically derived estimates based on model averaging. "

Comment: Line 291: How does this ozone metric correspond to the exposure metrics specified in each epidemiological study?

Response: SOMO35 metric is recommended by the EEA and also recommended in the latest WHO report reviewing the different ERFs. We have rephrased this part as follows (Line 358-360): "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: Line 292: Here (or elsewhere) it would be useful to provide the rationale for selecting these health endpoints. Citing back to WHO or US EPA documents or other systematic reviews would be helpful.

Response: We have now refereed to EEA and WHO reports in several parts of the manuscript (Lines XXX).

Comment: Line 297: It's really difficult to understand why YOLL are being divided by 10.6. Why not simply quantify counts of excess cases in the EVA tool?

Response: see comment to lines 303-321

Comment: Line 300: the selection of c-r functions greatly influences the health impact assessment, and so I'd recommend including this information directly in the manuscript rather than citing back to another paper. Likewise, what is the source of the baseline death and morbidity rates? At what spatial scale were these data available?

Response: We have not extended the section describing EVA substantially (Lines 326-464).

Comment: Lines 303-321: I'd suggest providing a clearer rationale for valuing years of life lost rather than counts of excess death.

Response: government agencies in Europe, including the European Commission, apply a methodology for costing of air pollution that is based on accounting for lost life years, rather than for entire statistical lives as is customary in USA. Whereas the average traffic victim, for instance, is mid-aged and likely to lose about 35-40 years of life expectancy, pollution victims are believed to suffer significantly smaller losses of years (EAHEAP, 1999:64; Friedrich and Bickel, 2001). To avoid overstating the benefits of air pollution control, these are treated as proportional to the number of life years lost.

The average loss of lifeyears per victim has previously been assessed to 10.6 (calculation method explained in Andersen 2017).

Comment: Line 314: Please provide a citation to support this claim.

Response: OECD, 2016 reference is now added to the text (Line 440)

Comment: Line 316: Did you consider adjusting the WTP to account for changes in income over time (i.e. income elasticity)?

Response: Indeed- the costs reported are the net present costs related to mortality and morbidity, and WTP is expected to increase with increasing incomes in the future; however this future stream of WTP needs to be discounted back into net present values. It has been customary in EU studies to apply an income elasticity of 1.

Comment: Line 320: Why adjust the WTP using a PPP when you can just apply a U.S. specific value?

Response: We have now extended this section (Lines 448-464). Cost-benefit analysis in USA relating to air pollution proceeds from a standard approach whereby abatement measures preventing premature mortality are considered according to the number of statistical fatalities avoided, which are appreciated according to the value of statistical life (VSL) (presently USD 7.4 million). In contrast, and following recommendations from the UK working group on Economic Appraisal of the Health Effects of Air Pollution (EAHEAP, 1999), focus in EU has been on the possible changes in average life expectancy resulting from air pollution. In EU the specific number of life years lost as a result of changes in air pollution exposures are estimated based on lifetable methodology, and monetized with Value-Of-Life-Year (VOLY) unit estimates (Holland et al. 1999; Leksell and Rabl 2001). The theoretical basis is a life-time consumption model according to which the preferences for risk reduction will reflect expected utility of consumption for remaining life years (Hammitt 2007; OECD 2006:204). The much lower VSL values customary in Europe (presently €2.2 million) add decisively to the differences, as VOLY is deducted from this value. By using a common valuation framework according the EU approach we allow for direct comparisons of the monetary results. It follows from OECD recommendations (2012) to correct with PPP when doing such benefit transfer.

Comment: Line 394-402: Please report the currency year.

Response: The currency year is 2013 (Line 464).

Comment: Line 418: Did you consider reporting population-normalized results (e.g. deaths per 100k)?

Response: such a figure is embedded in the specific exposure-response function for mortality, which was derived from lifetable analysis, however providing lost life-years per 100k

Comment: Line 434: Can you clarify what a health impact index is?

Response: We have now rephrased this paragraph (Lines 636-643): “Results show that for the particular input (gridded air pollutant concentrations from individual model)-output (each health outcome) configuration, the $PM_{2.5}$ drives the variability of the different health impact and that at least 81% of the variation of the health impacts are explained by sole variations in the pollutants (i.e. without interactions: Table S3). Table S1 also shows that the most important contribution to the health impacts is from $PM_{2.5}$, followed by CO and O_3 (with much smaller influence though). The impact of perturbing $PM_{2.5}$ by a fixed fraction of its standard deviation on the health impact is roughly double compared to CO and O_3.”

Comment: Table 2: The nomenclature is a little misleading. In a health impact function, effect coefficients are exponentiated and multiplied against an air quality change and then against baseline incidence rate and the population exposed. However, the effect coefficient is written as “x cases/ugm3”. This is not correct.

Response: In EVA, we use linear functions for the ERFs. We have now added the following section (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: Table 2: Several of the endpoints list multiple studies. Were these pooled in some way?

Response: Each of the morbidity effects refer to one study each.

Comment: Tables 3-4: Please include 95% confidence intervals

Response: We have moved the big tables into the supplementary material and made a new Table 3, which summarizes the mean results from the different ensemble approaches. Along with the mean of all individual pollutant estimates (denoted as MM_{mi} in the manuscript), we have now added the standard deviations. EVA model implements the ERF functions as linear equations and the 95% CI are not taken into account presently. We agree with the reviewer that it is important to provide these numbers, however the present study employs a frozen version of the model, where the aim is not focusing on further development of the model. We continue to further develop the model on many aspects and this comment will also be taken into account.