Response to Reviewer #2

We thank Reviewer #2 for evaluating our manuscript. Below, we list our responses to each comment (in blue). We first note that all analyses have been updated based on current model availability. This includes the addition of more realizations and/or climate variables to the models previously used (e.g., we now have MIROC daily data and 2 additional CESM2-WACCM simulations, etc.). We have also added two additional models to the analysis: NorESM2-LM and UKESM1-0-LL. Our overall results and conclusions remain unchanged.

Reviewer #2

Allen et al. use model output from the AerChemMIP intercomparison project to evaluate 2015-2055 changes in climate variables associated with two future air quality control scenarios. By comparing a "weak" policy scenario to a "strong" policy scenario, they show increasing trends in temperature and precipitation over the period that are driven by near-term climate forcers (ozone and aerosols), suggesting a climate penalty associated with air quality improvements. The manuscript is generally well written and well structured and makes good use of the AerChemMIP simulations. It addresses an important question that is well suited to the scope of ACP. I do have a few concerns about the statistics and a few more minor comments and suggestions, discussed below.

GENERAL COMMENTS

1. The trends have been calculated using least squares regression. There is very little information on exactly how that was implemented, so from my reading it does not appear that this is a weighted least squares regression, or that the uncertainties have been accounted for in any other way. This is concerning because, looking at Fig. 3 for example, there is a large amount of variability in the individual models that are used to construct the multi-model means. I am not convinced by the robustness of some of the reported trends in the multi-model mean, or that they are truly statistically significant as stated. The multi-model mean trend calculations should be performed using a method that accounts for variability/uncertainty in the mean (e.g., weighted least squares, but there are other options) before the paper is publishable in ACP. In addition, some discussion of the method used and the influence of the variability/uncertainty is warranted.

More information on the statistics are included. Models with multiple realizations are first averaged to form the model mean. Individual model mean trends are calculated using least squares regression, and the corresponding trend significance is based on a two-tailed Student's t-test, where the null hypothesis of a zero regression slope is evaluated. Autocorrelation of the time series is also accounted for by using the effective sample size: $n\times(1-r_1)/(1+r_1)$, where n is the number of years and r_1 is the lag-1 autocorrelation coefficient.

MMM trends and their significance are estimated in two ways, and both methods yield similar results. In the first approach, the overall multi-model mean time series is calculated as the mean over each model mean (i.e., each model has the same weight), and a similar procedure as above is used to determine the significance of the multi-model mean trend. However, we now use a

weighted least squares regression (as suggested by the Reviewer). Each value in the multi-model mean time series is weighted by $\frac{1}{\sigma_m^2}$, where σ_m is the standard deviation across models.

We note that this does not change any of our results. For example, the global annual mean multi-model surface temperature trend changes from 0.062 without weighting to 0.060 K/decade with weighting, and both are significant at the 99% confidence level. Weighting the regression introduces negligible changes in our other climate and air quality trends.

In Figure 3, there was actually quite a bit of similarity in individual global mean model trends. All but one individual model surface temperature trend in Figure 3 was significant (MIROC6 the lone exception). Furthermore, all individual model precipitation, Hottest Day, PM2.5 an Ozone global mean trends were significant. Weaker results existed for the Wettest Day, and in particular, Consecutive Dry Days. It is therefore not surprising the multi-model mean trends were also significant. Using the weighted regression approach, we get similar conclusions.

To summarize, we now use a weighted least squares regression for the multi-model mean trends, and we get similar results as before.

In addition to estimating the magnitude and significance of the multi-model mean trend as just described, we also evaluate the multi-model mean trend and its significance relative to the individual model mean trends (e.g., Figure 5). Here, the MMM trend is estimated as the average of each model mean trend, and its uncertainty is estimated as plus/minus twice the standard error (i.e., the 95% confidence interval). This is calculated as: $2\times\sigma/\sqrt{nm}$, where σ is the standard deviation of model mean trends and *nm* is the number of models. If this confidence interval does not include zero, then the multi-model mean trend is significant at the 95% confidence level.

The corresponding 95% confidence intervals are now included in each of the global time series plots. As is the R^2 value of the multi-model mean trend.

We have also added the multi-model global mean trend (and others, including NH mid-latitudes, Tropics, SH mid-latitudes) and its uncertainty to the bar graphs in Figure 5.

2. For the global trends in climate variables, it would help to contextualise the values associated with NTCFs by also providing the trends from the two individual scenarios(or at least from the one with weak NTCF control, as the other can be determined from the difference trends provided). Without this, it's hard to tell how important the NTCF climate penalty is. I note that this is done in the figures for the regional trends, but not for the global trends. I would strongly encourage the authors to add these in some form (for example, a figure in the SI equivalent to Fig. 3).

These have been added to Figure 5. The last set of bars (labeled "GL") now show the global mean trends for SSP3-7.0, SSP3-7.0-lowNTCF and their difference. Also included is the corresponding land (labeled "Ld") surface values (which were previously included). We have also added additional trends over various latitude bands.

3. The manuscript is very well structured and quite well written, but the heavy use of acronyms and technical identifiers (e.g., SSP3-7.0-lowNTCF, lowAERO3, etc.) makes it harder to read & follow than it needs to be. I would encourage the authors to simplify this wherever possible and then use a consistent, easy to interpret nomenclature throughout. For example, frequently the two scenarios are referred to as strong and weak air quality control, and these are much easier to interpret than SSP3-7.0 andSSP3-7.0-lowNTCF. I would suggest strong and weak air quality control could replaceSSP3-7.0 and SSP3-7.0-lowNTCF everywhere, in particular in figure legends and captions where the reader may not be referring back to the text. Similarly, NTCF mitigation is easier to interpret than SSP3-7.0-lowNTCFSSP3-7.0.

We have removed some of the acronyms. In particular, we now use strong and weak air quality control, as well as NTCF mitigation. We also use more straightforward acronyms for the two model subsets, Aer and Aer+O3, as suggested by Reviewer #1.

4. The manuscript cites a lot of "in prep" and "submitted" papers. In most cases, these are cited as part of long lists of other references, so they aren't really needed to make the points. If these are not at least in ACPD by the time of publication, they should not be included in the citation lists (except in cases where they are the only publications available to back-up the point).

All references have been updated.

SPECIFIC COMMENTS

L30: Does the net radiative effect here refer just to OC or to BC+OC? Please rephrase to clarify.

This statement has been clarified. This is the best estimate of net industrial-era climate forcing by all short-lived species from black-carbon-rich sources.

L59: Is this newer estimate of mortality for all air pollution or outdoor ambient only? Please rephrase to clarify.

This is for all air pollution. Clarified.

L90-108: This information would benefit from being summarised in a table listing the scenarios and some of the relevant information (e.g. air quality controls weak/strong,ozone and aerosols high/low, CH4 high/high, etc.) to make it easier for the reader to synthesise.

This information has been moved to this section ("Future Scenarios"), including Figures 1 and 2, which show the global evolution of emission species and the regional trends. Since only two scenarios are addressed in this manuscript, we only show results from SSP3-7.0 and SSP3-7.0-lowNTCF.

L120-122: I find this a bit confusing. What is the difference between CESM2 and CESM2-WACCM in this case? Is it the aerosol treatment? And if they are basically the same model, is it fair to include them as two separate data points in the multi-model means?

We have removed CESM2 from the analysis.

L141-144: So nitrate aerosol was not included in PM2.5 at all, even for the models that do include it? It would be nice to see how much uncertainty this adds, given nitrate can be an important component of aerosol loading in some regions. I'd suggest adding a version of the PM2.5 figures including nitrate to the supplement, and a brief discussion of the impacts of excluding nitrate either in the main text or in the supplement.

Only one model includes nitrate aerosol data (GFDL-ESM4). Globally (over land only), nitrate decreases by -0.0396 (-0.1165) μg/m^3. These trends are 17 and 20% of the magnitude of the corresponding PM2.5 trends. GFDL-ESM4 also archives ammonium. Globally (over land only), ammonium decreases by -0.0487 (-0.1168) μg/m^3. These trends are 21 and 20% of the magnitude of the corresponding PM2.5 trends. Thus, excluding nitrate and ammonium in GFDL-ESM4 leads to ~40% underestimation of the global PM2.5 trend.

CESM2-WACCM also archives ammonium. Here, however, the global (land) trends are much smaller at -0.00329 and -0.0081 μg/m^3, which leads to ~1% underestimation of the global PM2.5 trend.

This has been added to the revision, as have supplementary figures that show the spatial trend maps for nitrate and ammonium. We have also added a discussion and supplementary figures that compare archived versus estimated PM2.5 trends in 4 models (those 4 that included archived PM2.5).

L156: Is there a reference for these ground-based observations? Or is this the same GASSP observations mentioned above? If the latter, please state explicitly in the text.

This is referring to GASSP. Fixed.

L172-L180: This is confusing when paired with the figure. It is completely legitimate to not include the differences in CH4 pathways for this work, but anyone skimming quickly and focusing on the figures will miss that point. In my opinion, Figure 1 should only show what was used in this work, not scenarios that are not used here. I strongly encourage the authors to remove the SSP3-7.0-lowNTCF (right?) and difference lines from Figure 1. The comparison between the scenarios can be moved to the supplement if the authors feel it is important to include.

The SSP3-7.0-lowNTCF and difference CH4 data have been deleted from Figure 1 (same for Figure 2). We have also removed SSP3-7.0-lowNTCF CO2 from these figures. Both sets of AerChemMIP simulations use the same CO2 and CH4 data, based on SSP3-7.0.

L181-187: Similarly, I don't think this discussion belongs here. It is the first section of the results, yet it is mostly discussing what is not done in this work. I would suggest this could be

removed entirely, or moved to the supplement or to the conclusions as part of a discussion of what future work should be done to build on what the authors have done here.

This discussion has been removed.

Sect. 3.2 and Figs 3-4: Generally speaking, is the changes in atmospheric composition (aerosols and ozone) that are driving the changes in climate. Thus it seems a bit odd to show and discuss the changes in climate variables BEFORE the changes in composition (ozone, PM2.5). I would suggest restructuring such that Fig.3 comes before Fig. 4, with the text order changed to match. (I note this is already the order used in the abstract and conclusions).

Re-ordered according to the reviewer's suggestion.

L204-205: Unless you rename & define the scenarios in the methods as discussed above, please clarify how "under NTCF mitigation" is defined here (I understand that itis the difference between the two scenarios, but that wasn't clear to me on first read).

Scenarios have been defined according to the reviewer's suggestion. SSP3-7.0 is referred to as weak air quality control and SSP3-7.0-lowNTCF is referred to as strong air quality control. Their difference (strong minus weak air quality control) is referred to as NTCF mitigation.

L211-218, L223-228 (and elsewhere): Much is made of the difference between the lowAER and lowAERO3 outcomes. Given that one of these only includes 3 models and the manuscript states explicitly that the difference is not significant, it is not justifiable to be interpreting this as a result. This appears to me to be over-interpretation of noise, and I would suggest this discussion be removed before publication in ACP.

We agree, and this statement has been modified. It is still interesting that similar (i.e., non-significant differences) global mean surface air temperature trends occur in Aer and Aer+O3 models. We acknowledge that this could be due to several factors, but one interpretation is weak surface cooling due to reductions in ozone.

L233: This land-only result appears to be insignificant for 75% of the models (including those that show increases) and so this statement should be removed or qualified.

Sentence has been deleted.

L238: CDD does not show a statistically significant increase in the overall MMM (or in the subset MMM or in any of the individual models bar one) – therefore should be removed from this sentence.

Deleted.

L255-256: Is the land-only warming pattern shown anywhere? Is the land-only warming weaker or stronger than the overall warming? If there is a difference, it would be useful to see an

equivalent figure in the supplement. (And if there is not a difference,it's not clear why this is discussed separately.)

Table 1 shows that the land warming is stronger than that over both land and ocean. This has been clarified. Surface temperature trend patterns are included in the Supplement.

L264: "...forcing and response do not need to occur in the same regions." Can this be explained a bit more?

A sentence has been added.

L269-271: Do I understand Fig 6 bottom panel correctly that models don't agree about this feature? If so that would be worth stating in this discussion

Figure 6 bottom panels show the percentage of models that agree on the sign of the trend. Red colors indicate model agreement on a positive trend; blue colors indicate model agreement on a negative trend. White areas indicate lack of agreement on the sign of the trend. The caption has been clarified. About 70% of the models agree that the North Atlantic cools.

L307: For a discussion of seasonal patterns to make sense, consideration should be given to the different seasonalities of the two hemispheres. Figure 7 should either be separated or at least ordered/demarcated by hemisphere – I'd suggest NH extratropics, tropics, and SH tropics.

Figure 7 shows seasonal trends for each of our 12 world regions. Thus, this figure is already broken down into regional demarcations consistent with seasonality. Nonetheless, we have also added trends for several latitude bands, including those requested.

L370: Why is one model listed explicitly when all (including that model) are available from the same location? Also please spell out ESGF here and provide a link or doi.

Reference to GFDL deleted. ESGF spelled out, and a link is provided.

Figs 2, 5,7: regional legend labels on x-axis are impossible to read because they are so small. Perhaps give each region a number instead? Or include some other sort of key to make this clearer?

We have modified the x-axis on these figures. A key is now used.

Fig 3 caption (and elsewhere): Does "hottest day" refer to "surface temperature on hottest day"? Similarly for wettest day? Please clarify somewhere.

Extreme weather indices are defined in the Methodology section.

We also analyze climate extremes including the hottest day (monthly maximum value of daily maximum surface temperature), wettest day (monthly maximum 1-day surface precipitation) and consecutive dry days (CDD), defined as the maximum annual number of consecutive days with

surface precipitation less than 1 mm/day. We focus on these three extreme indices since they are frequently used metrics for temperature and precipitation extremes. Prior observational analyses have shown significant increases (decreases) in the hottest and wettest day (CDD) over the latter half of the 20th century (Donat et al, 2013a,b). Climate extremes are based on daily data, and are calculated at each grid box and then spatially averaged.

Fig 3 caption: It seems the thin coloured lines show the trends for the individual models, but this has not been explicitly stated in the caption. Please update caption to clarify.

Thin (and non-black lines) show individual model mean trends. Line colors are denoted by the legend. We have also added this to the caption. Same for Figure 4.

Figs 3, 4, 5: why are different units used for the trends in the precip variables (mm/dayvs. %) in the global and regional trend figures? Same question for PM2.5 and O3. Can these be standardised to more easily compare?

Sure. We no longer use percent changes.

Fig 6d-f: these plots are not currently discussed in the text and therefore should per-haps move to the supplement (or be mentioned in the text)

A sentence pertaining to these panels has been added.

Table 1: I found this table hard to understand while trying to refer to it while reading the text. A few suggestions to improve the clarity. (1) Add lowAER and lowAERO3 identifiers above the list of relevant models in each sub-section so it's easy to see which group is which. (2) If text and figures are re-ordered as suggested above, move PM2.5 and O3 columns to be left-most, followed by the climate variables. (3) Move the three "MMM total" lines either to a separate part of the table or (preferably) to a new table altogether as the numbers aren't comparable to the lines above/below which makes it difficult to interpret (and already a lot to interpret in the table!). (4) For the lowAER models' O3 response, replace 0.0 with n/a since these values are not included in the Overall MMM calculation (as is, looks like the overall will be an average of thelowAERO3 values and three zeros).

Made all suggested modifications to Table 1.

TECHNICAL COMMENTS

L139: "were are" –> "we are"

L357: "complex" –> "complexity"

Fig 2 caption: "astriks" –> "asterisks"

All have been fixed.