Response to Reviewer #1

We thank Reviewer #1 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 #1

General:

Allen et al. introduce results of the AerChemMIP project on the impact of air quality measures on climate. This is a large exercise and certainly worth publishing. However, I think there are major shortcomings. The most apparent is the style. The paper is written as a report, stating what has been done and what is the outcome. While this is, of course, an essential part of a paper, it should contain much more. It is less written as a scientific paper that should motivate chosen assumptions, extract main new massages from results, and discuss uncertainties e.g. wrt. to the chosen assumptions. This is largely missing. For example the main message "Our findings suggest that future policies that aggressively target non-methane NTCF reductions will improve air quality, but will lead to additional surface warming" is shown in the end as being nothing new, but already covered by many other studies, as shown by the authors in line 345ff. So what is new? And this puts me actually in a difficult position, why should a paper be published which "just" confirms previous findings? I understand that IPCC deadlines have to be met, but more emphasis should be given to clearly describe what is new. More examples are given in the detailed comments below.

We have attempted to improve the writing, by placing more emphasis on motivation and assumptions. Although our main results support prior studies, given the sophistication of the models used here, as well as the relatively large number of models, we suggest that this work is the most comprehensive analysis on this topic to date. Overall, the structure of the manuscript is similar to the original submission. Reviewer #2 states: "…the manuscript is generally well written and well-structured and makes good use of the AerChemMIP simulations".

Major comments in addition to the writing style:
1) Structure: The method section is too short:

- More information on statistics should be given (see details below); Please explain why the multi-model trends are significant, although individual model trends are not. What trend model has been used? What exactly is tested?

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 Reviewer #2). 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.

- A motivation why exactly these climate/air quality/extreme indicators are chosen is missing;

As discussed in the Introduction, both PM2.5 and ozone are commonly used indicators of air quality. Both have been associated with adverse human health impacts. Surface temperature and precipitation are analyzed as these are arguably two of the most important climate variables. Changes in surface temperature are particularly relevant in the context of climate mitigation, as

the goal of the Paris Agreement is to keep the increase in global mean surface temperature to well below 2°C above preindustrial values. Precipitation, and fresh water resources in general, are important to both human society and ecosystems. Perhaps more important than changes in the mean of a climate variable are changes in its extremes. Heat waves, for example, are a major cause of weather related fatalities. We focus here on the hottest and wettest day, as well as consecutive dry days, as these are frequently used extreme temperature and precipitation indices (e.g., Donat et al., 2013a,b). Furthermore, prior observational analyses have shown significant changes in all three quantities over the latter half of the 20th century. This information has been added.

- Part 3.1 is actually input to the study and should be moved from the results part to the method part.

Moved to the Methods section.

2) Statistics: I have strong concerns how the statistics are interpreted. If a difference is not statistical significant, there is no basis in discussing them. Please remove all parts, which interpret statistically insignificant differences.

These have been removed.

3) Discussion: How important are the choices made in the assumption section?

We assume the "assumption section" pertains to the future emission pathways used here. We have added more information on the assumptions pertaining to the two future emissions pathways analyzed here. Basically, our analysis assumes that NTCF policies can be enacted in the absence of GHG related climate policies (e.g., SSP1's air pollutant legislation and technological progress can be achieved in the SSP3 world). Furthermore, our results likely represent an upper bound, since our baseline/reference scenario lacks climate policy and has the highest levels of NTCFs. This has been clarified in the Future Scenarios Section.

In the Conclusion Section, we discuss the implications of the assumptions made in the weak and strong air quality control pathways used in this analysis. It is not possible to formally quantify these assumptions, as different NTCF mitigation simulations were not performed by AerChemMIP. We have added clarification in the conclusions:

Our simulations, however, do not account for CO2 reductions, implying the importance of simultaneous reductions in both CO2 and NTCFs. We note that it is difficult to reduce only the NTCF emissions while keeping CO2 emissions fixed (since there are co-emitted species, including SO2). If CO2 emissions are simultaneously reduced along with NTCFs, then the increase in global surface temperature and precipitation found here will be muted.

Detailed comments:

Abstract: "How future policies affect the abundance of NTCFs and their impact on climate and air quality remains uncertain." I am wondering whether this could be misunderstood in a way that for a given measure the impact remains uncertain. Most of the uncertainty comes from the uncertainty what measures will be taken, right?

Future climate and air quality are uncertain for two reasons. There is uncertainty due to the emissions pathway, and there is uncertainty in the corresponding climate response. Past IPCC reports have shown that both uncertainties are approximately of the same magnitude in the context of climate change. The latter uncertainty is due to uncertainty in climate sensitivity (e.g., 1.5-4.5 K per 2xCO2). As an aside, CMIP6 models tend to have a higher climate sensitivity than CMIP5 models, which has been related to clouds (e.g., Zelinka et al., 2020). Nonetheless, we have attempted to clarify this sentence, since the larger uncertainty for a given pathway is the climate response.

l13 "similar increases" what means similar here? Can a extreme weather index be similar to a temperature increase of 0.24 K? or is even 0.34 K similar to 1.1%. Please specify.

Re-worded.

l16 "ozone reductions.": I think it would be helpful to include half a sentence explaining the relation between aerosols and ozone.

We have added information in the Introduction. "…reductions in some precursor gases such as NOx and VOCs impact both ozone and aerosols (and perhaps CH4). Reductions in NOx, for example, will promote cooling due to reduced tropospheric ozone, but the impact on CH4 lifetime and aerosol formation will likely promote overall warming"

l20-21: I think the definition in Myhre et al 2013 is "We define 'near-term climate forcers'(NTCFs) as those compounds whose impact on climate occurs primarily within the first decade after their emission." It reads a little bit different from "that impact climate on relatively short time scales, typically within a few weeks to a decade after emission". Climate is defined on decadal timescales. To relate climate change to to weeks sounds weird. Concentration changes and RF can quickly react, but you started to discuss climatological changes in temperature and rain rates and those do not occur on weekly timescales. Please adapt the text.

We have adopted the reviewer's verbiage.

l28 should it be "-2.0 to−0.4" ?

Re-ordered.

l34 shouldn't methane be mentioned here as well, since it is a precursor for ozone? I think you are referring to table 8.6 in Myhre et al. 2013. Their tropospheric ozone area total ozone change and include effects from methane emissions.

Methane added here.

l34-37: Here you change from a concentration perspective (ozone) to an emission perspective (methane). Please clarify this, otherwise it seems to be inconsistent and double counting methane ozone effects. Especially the wording "Similarly," should be revised, since the view is exactly not similar.

Modified to concentration perspective.

l42-44 please clarify the sentence. How can a change in radiation, i.e. in W/m2, be balanced by evaporation in units m/s.

Changed evaporation to latent cooling.

l62 please clarify what you mean with "rapid". See also discussion above.

Clarified. Added “decadal”.

l91 You mean the scenarios you are employing. ...

Added “Used here”.

Section 2.1: I think it would be nice to have a motivation included. Currently, it reads like a report or namelist setting. Why is the reference without climate policy? etc. this should be motivated.

We have added motivation, assumptions, and clarity. Our analysis assumes that NTCF policies can be enacted in the absence of GHG related climate policies (e.g., SSP1’s air pollutant legislation and technological progress can be achieved in the SSP3 world). Furthermore, our results likely represent an upper bound, since our baseline/reference scenario lacks climate policy and has the highest levels of NTCFs (i.e., to detect the largest signal, the reference is without climate policy).

Please also include a table showing the changes in relevant emissions, such as aerosol compounds and ozone precursors for some well-chosen times, e.g. 2015, 2035, 2055;or decadal? I think it is important to see the changes.

Figure 2 already shows changes in emissions by region, as well as over land.

l120 I find the abbreviation misleading. "lowAER" and "lowAERO3" are model group names. "low", however, is not referring to the models, but to the scenario, right? and at some point I though "AERO3" is the "AEROsol Group 3" and not aerosol-ozone. What about "Only-Aer" and "Aer-O3"; or "Aer+O3" ?

Changed abbreviations to “Aer” and “Aer+O3”

l135 Please include what kind of linear model you are using y(t)=a+bt+err or y(t)=b(t-2035)+err ? Are you fitting one or two parameters? Often as a measure for the fitting quality the R^2 value or adjusted R^2 value is used. Why not here? I do not understand how the trend is tested. Are the individual model results fitted and then tested whether the mean trend is representing the range of models correctly? (At least the caption of Figure 7 might indicate something like this). How the statistics are treated is very important for the interpretation. Please include a thorough discussion here.

Our response to the concern over statistics is located above.

l159 Also here a motivation is missing. I understand that extreme values are important. But why is the max temperature chosen? Isn't that a statistically very difficult quantity, even among extreme value statistics? Why not using number of hot days, i.e.over 30◦C? This also concentrates on extremes, but includes a whole tail of a pdf (or estimated pdf).

The Hottest Day ("TXx") is chosen as it is a commonly used extreme temperature measure. See for example, Donat et al. (2013 a,b). As with other extreme temperature indices, significant increases in TXx were found (1951-2011) in two different data sets, GHCNDEX and HadGHCND. We also find significant TXx trends in the simulations analyzed here.

l162 please also add the respective time frame. Are you averaging over 10 or 30 years?

We are not sure what time frame the reviewer refers to. L162 states "Climate extremes are calculated at each grid box and then spatially averaged." There is no decadal averaging.

The hottest day (monthly maximum value of daily maximum temperature) and the wettest day (monthly maximum 1-day precipitation) are estimated for each month, and then averaged to obtain annual means. Consecutive dry days (CDD), defined as the maximum annual number of consecutive days with precipitation <1 mm/day, are estimated for each year.

l165ff: I would have expected this part in the scenario section. Please consider to move it there, since this is not a result from your paper, right? And then ignore my comment on the table (see above) ...

Moved to scenario section.

l167: Why is there a CO2 emission change at all, if you are considering NTCF changes only? Please explain. I don't think that this is a problem, but currently and certainly it confuses me.

Yes, there are small differences in CO2 emissions between the two scenarios. Methane reductions generate emissions abatement costs, which changes industrial outputs in all productin sectors and household consumption (Gidden et al., 2019). Energy consumption and CO2

emissions in all sectors are thus affected, which causes small differences between SSP3-7.0 and SSP3-LowNTCF.

However, AerChemMIP simulations use the same CO2 emissions, based on SSP3-7.0 (as with methane). This has been clarified in the revision. We have also removed the SSP3-7.0-lowNTCF CO2 (and CH4) emissions from the plots, to avoid unnecessary confusion.

l192: Why are you discussing methane emission changes, if those are not relevant?

Deleted discussion on methane emission changes.

Figure 2: Trends are calculated as (2055-2015)/4 or with the regression method discussed in Section 2?

Emission trends are calculated using the same method as above. Trends are based on a least squares regression, with significance based on a two-tailed Student's t-test. We note that the emissions data is decadal after 2015, with monthly values for the year 2015, 2020, 2030, 2040, 2050, 2060, etc. We estimate the emissions in 2055 as the mean of the emissions in 2050 and 2060 at each grid box. We have added this information.

l 205: Please comment if the trends of the individual models are statistical significant. I miss a mathematical/statistical explanation in combination with a motivation why to test the multi-model mean and not, whether mean trend is significant wth respect to the variation in trends of the individual models.

Table 1 lists the trend (and whether it is significant at the 95% confidence level) for each model. All but one model yields a significant global mean increase in surface temperature. This has been clarified.

We initially did not evaluate the significance of the multi-model mean trend, relative to the individual model trends, for the global mean quantities. We did this for the regional trends (e.g., Figure 5). We do note that Figure 5 included land only averages. Nonetheless, we have now performed this analysis for the global mean quantities. The 95% confidence interval is now included in the global time series plots. We have also added the multi-model global mean trend and its uncertainty to the bar graphs in Figure 5 (and additional latitudinal bands).

l 206: For the regional trend an uncertainty range is given. Why not here?

Yes, we have now added this analysis for the global trends. The corresponding 95% confidence intervals are now included in each of the global time series plots. We have also added the multi-model global mean trend and its uncertainty to the bar graphs in Figure 5.

l213-l214: If a result is not statistical significant, there is no point in interpreting the result. Please delete the sentence.

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.

l223: Keep in mind that the change was not statistical significant; so the results may not be inconsistent, but only noise. Please revise the discussion, based on what is inconsistent on a statistical significant basis.

Discussion edited and revised. As with surface temperature, non-significant differences between Aer and Aer+O3 ERF trends exist.

l 246: "Slightly larger (but not statistically significant) "; if not statistically significant, then they are not slightly larger! Please respect the statistics.

Edited. Similar increases occur in both Aer and Aer+O3 models.

l263-264: Please rephrase the sentence. I agree with the content, however, the formulation, starting with "however" suggests that there is either a shortcoming or something unexpected, etc. As the authors state this is by no means a surprise nor limitation of the aforementioned.

Deleted "However"

l262 I think somewhere it should made clear that a part of the warming is a reduced cooling from SO2 reduction and O3 reductions, right?

Yes, based on the information presented in the Introduction (e.g., radiative forcing), SO2 reductions will warm. But O3 reductions will cool. This has been clarified.

From the Introduction: Thus, reductions in some NTCFs, including non-absorbing aerosols, will warm the climate system, whereas reductions in other NTCFs, including absorbing aerosols, tropospheric ozone, and methane will cool the climate system.

l284 Is there some relation to the monsoon tipping points?

L284 states: "Furthermore, in agreement with prior studies, precipitation increases in several monsoon regions, including east Africa, south Asia, and east Asia." Thus, unlike the buildup of aerosols over the 20th century, future NTCF mitigation and continued increases in GHGs will likely accelerate the monsoons. Not exactly sure what the Reviewer wants us to change here.

Section 3.4: What about MAM/SON? Discussed are winter/summer differences. However, "Seasons" would imply more than that. I suggest to, at least, mention a general trend for MAM/SON and add the same figures in a supplement.

Added general trends in MAM/SON seasons. Added MAM/SON plots to supplement.

l339-342: This is important: see also above. If the difference is not statistical significant, there is no point in discussing or even highlighting it in the summary. Please remove this part!

We have rephrased. The lack of significant trend differences in Aer and Aer+O3 models is interesting. We acknowledge that this could be related to several factors. But one possible interpretation is weak surface cooling due to reductions in ozone. We feel as if our ability to compare Aer and Aer+O3 models is one of the novelties of this study. But again, we acknowledge this comes with caveats.

l359: It might be worth mentioning reduced warming, i.e. a net cooling. To avoid confusion about weather CH4 itself has a cooling contribution.

Added “net cooling” here.