Reviewer #1:

I thank the reviewer for thoughtful and constructive comments.

The author aims to constrain both the climate sensitivity and the aerosol-induced cooling (between pre-industrial to present-day) from the time evolution of the global mean surface temperature. The constraint comes from the decoupling between the greenhouse gas and aerosol forcings that occurred between 1980 and 2000 and its impact on the historical temperature record. The author finds a smaller aerosol cooling effect than predicted in the models but also a smaller aerosol cooling effect than generally found in other detection and attribution studies. While this manuscript is complementary to previous studies and it represents an interesting contribution to the literature, I doubt it is the end of the story.

I doubt that too.

Furthermore I have a number of major comments that would require clarifications and most likely further analysis. Overall I recommend major revisions to the manuscript before it can be considered for publication in Atmospheric Chemistry and Physics.

Major comments

1/ The author should do a much better job at citing and describing previous work on the topic (e.g. doi: 10.1038/s41558-020-00965-9, 10.1175/JCLI-D-19-0091.1, 10.1038/ngeo2670 but I am sure there are other references as well). There is a large variety in approaches. Storelvmo et al. used a multi-variate fingerprint of the aerosol impact on the climate system. Charles et al. relied on time variations of the ocean heat content. Gillet et al. used not only the magnitude but also the pattern of the aerosol cooling to constrain the climate sensitivity and forcing. The uncertainties remain large but available studies generally find a larger contribution of the aerosols (relative to greenhouse gases) than found in this manuscript. For instance Gillett et al. (2021) concluded that "Greenhouse gases and aerosols contributed changes of 1.2 to 1.9 °C and −0.7 to −0.1 °C, respectively, and natural forcings contributed negligibly." The author should discuss or at least speculate why his conclusions are somewhat different from those of past studies.

Commented [OM1]: Gillett et al., NCC, 2021

Commented [OM2]: Elodie, C., Meyssignac, B., & Ribes, A. (2020). Observational constraint on greenhouse gas and aerosol contributions to global ocean heat content changes. *Journal of Climate, 33*(24), 10579-10591. doi:https://doi.org/10.1175/JCLI-D-19-0091.1

Commented [OM3]: Storelvmo, T., Leirvik, T., Lohmann, U. *et al.* Disentangling greenhouse warming and aerosol cooling to reveal Earth's climate sensitivity. *Nature Geosci* **9**, 286–289 (2016). https://doi.org/10.1038/ngeo2670

I have now strengthened the discussion of the literature. Previous authors including Gillett et al (2021) have found a highly uncertain aerosol-induced cooling that my results are generally consistent with. The main contribution of this paper is that I explicitly consider additivity and disqualify models that do not satisfy additivity. This results in less model disagreement. Previous studies (e.g. Gillett et al. 2021) have generally assumed additivity, have included all available models, and as a result have derived quite uncertain results. This is now discussed in some detail.

2/ The author compares the global surface air temperature (variable tas) from the models to the global-mean surface temperature from HadCRUT5 but fails to recognize that these are two different things. HadCRUT5 is actually a mix of SST over the ocean and surface air temperature over land. It should be noted that the trends in GMST and GSAT are somewhat different because of the contributions from regions of melting sea ice. Indeed GSAT increases a lot in places where sea ice melts while SST is hardly affected. Furthermore the kink in GMST and GSAT due to aerosols may not be synchronous. Both IPCC and Gillett et al (2021) moved away from GMST to embrace GSAT. I

strongly recommend that the author repeats his analysis with GSAT observations (such as provided by the Berkeley Earth project) to avoid the inconsistency in the way temperature trends are estimated between models and observations.

To improve consistency with HadCRUT5, I now use as the basis for comparison with HadCRUT5 an "amalgamated" temperature product which is SST over the ocean and SAT over land. (This is how HadCRUT5 is constructed.) The results only differ marginally from what I had originally. I understand that this is done because surface-air temperature is what land-based met stations routinely measure, whereas historic ship-based observations were usually of SST. Given the insensitivity of the results with regard to the choice of ts or tas, only HadCRUT5 is used here, and model data are now used in a way that is consistent with this choice.

3/ It is annoying and worrying that the total warming (from the sum of the three experiments) differs so much from the historical warming given that climate models are generally thought to be largely additive for small perturbations. Is the plot in Fig 1b for the ensemble mean?

Indeed these numbers reflect the ensemble means. Actually several small changes to the regression model (namely the introduction of an "intercept" δ in the new Eq. 2, which accounts for small problems with the normalization of the temperature datasets and makes the regression symmetric, also smoothing of the hist-ghg and hist-aer ensemble means) for most models improves this situation. Also I now re-interpret the terms $\sim T_{haer}$ to stand for the total influence of near-term climate forcers (NTCFs, i.e. aerosols and ozone). This means the expected value for β2 is now < 1. I now discuss that eight models satisfy additivity (i.e. (a) $|\alpha_2 - 1| <= 0.2$ and (b) $0.6 <= \beta_2 <= 1.1$). In the calculation of the corrected ECS, TCR, and NTCF-induced cooling only those models are used that satisfy both (a) and (b) are considered. Selection of only those models yields substantially smaller statistical uncertainties for the resultant best-estimate correction factors because some models yielding relatively extreme correction factors are then systematically excluded. I think exclusion of the non-additive models is the main reason for the reduced model uncertainty found here, relative to other papers. However, the flip side of the coin is that now this is based on a relatively small number of models (i.e. 8). I think this is the smaller price to pay, relative to including models that are unsuitable for this attribution study because they are not additive.

How large is the ensemble?

Ensemble sizes are given in table 1.

How significant is the non-additivity?

For some models the non-additivity is substantial. Those models are now not considered in calculations of best-estimate correction factors, TCRs, ECSs, and adjusted NTCF-induced cooling. More models satisfy additivity for the GHG influence than the aerosol influence. I have therefore introduced a more stringent condition ($|\alpha_2 - 1| < 0.2$) for the GHG additivity (which removed 5 models) than for the aerosol influence ($0.6 <= \beta 2 <= 1.1$). This removes 8 models. 8 models remain that satisfy both criteria. Consideration of models that satisfy additivity considerably reduces uncertainties for all multi-model means calculated here. Figure 1b indicates that the 8 models retained for the analysis all exhibit a relatively good correspondence between simulated "historical" warming and warming summed up over the three sensitivity experiments. The eight excluded models comprise four (GFDL-ESM4, MRI-ESM2, CanESM5, NorESM2-LR) where this total warming is outside the 80-120% envelope of simulated historical warming.

Can it be explained by natural variability or missing terms in the total warming?

The main missing term is the ozone forcing. This is now considered in two ways: Firstly, I now provide an emergent-constraint analysis of warming due to the combination of aerosols and ozone, however based only on the hist-aer experiment. The underlying assumption here is that the warming due to ozone is proportional to that due to aerosols. According to IPCC (figure 7.8) this is a pretty good assumption. Throughout 1850-2019 ozone has consistently offset ~1/3 of the cooling due to aerosols, a little more in the beginning (when both forcings were small) and in the end when aerosols are decreasing). This motivates that I now have an expected value of β_2 of ~0.7.

In addition, I have repeated the analysis for the 5 models that have conducted the hist-totalO3 experiment. For CanESM5, the analysis confirms that this model is substantially non-additive. For MIROC6, β_2 goes from 0.76 to 1.01 (in agreement with the above expectation), but in this model, the ozone forcing hardly produces any sustained warming, so this may be fortuitous. In HadGEM3 and GISS-E2-1-G, the coefficients remain largely unchanged. MPI-ESM1-2-LR is non-additive when ozone forcing is included, because warming due to ozone is very nearly of the same size and directly opposed to aerosol-induced cooling in this model, making them impossible to separate.

So with the presently available hist-totalO3 simulations, a robust analysis of the influence of ozone on global-mean temperature remains very difficult.

And if so what is the impact on the author's analysis? If the departure of α_2 and β_2 from unity is due to natural climate variability (which I suspect), then I do not see the rationale for normalizing α_1 and β_1 with α_2 and β_2 as done in Eq. 3. If the departure of α_2 and β_2 from unity is instead due to missing forcing terms, then should the missing term not be inserted in Eq 1? Fig 4 shows that the departure of the black ellipses from the (1,1) point is quite generalized. If it was due to climatological noise, then should we not expect values that are both smaller and larger than 1? Here we have mostly values smaller than 1 (especially for β_2). Such low values question the validity of the framework. I note that the author discusses this issue in the conclusion section (line 227ff), however this comes far too late in the manuscript. The fact that $\beta_2 < 0.5$ for several model does not only "question the suitability of these models for attribution", it questions the suitability of the method. I would strongly recommend the author to investigate further the reason for non-unity values of α_2 and β_2 and its impact on the analysis. This is a show-stopper in my opinion.

As noted, there is no conclusive evidence that the non-additivity is due to missing forcing terms. I cannot straightforwardly expand the regression model because other experiments in DAMIP are tier-2, and few models have conducted these simulations.

I agree with the reviewer that the best course of action w.r.t. non-additivity is not to include such models in calculations of multi-model averages. This is what I have done in the revised version of the manuscript.

The main impact is the introduction of a straightforward criterion to distinguish between suitable and unsuitable models. Based on this criterion uncertainty ranges for both the GHG-induced warming (ECS, TCR) as well as the aerosol-induced cooling are reduced versus other assessments. Also the best-estimate aerosol-induced cooling is smaller than found in other studies, though well within their uncertainty ranges.

4/ How do uncertainties in T_{obs}, not accounted for in the current analysis, affect the results?

We have replaced HadCRUT5 with Berkeley Earth, an alternative global temperature dataset (the variant where SSTs over sea ice are inferred from open water around sea ice). Berkeley Earth uses the same HadSST4 dataset for sea-surface temperature as HadCRUT5. Hence there are no qualitative

differences in the results, indicating that model differences for outweigh any observational issues for this analysis.

5/ ECS computed from abrupt 4xCO2 simulations is generally larger than when computed from abrupt 2xCO2 simulations. It is also an imperfect predictor of the warming predicted in transient simulations. It is not clear to me why the author considers ECS rather than TCR as a predictor of GHG or total historical warming.

It is imperfect but certainly useful, see figure 1b. The ECS values quoted here are derived from 4xCO2 simulations. While there are differences between the two types of experiments used to quantify ECS, the results should be highly correlated, and most ECS values are derived from 4xCO2 experiments that are part of the ubiquitous DECK family of experiments. We now treat the TCR in the same way as the ECS. Interestingly, our result for the ECS (3.4K) is larger than the IPCC best estimate (3K), while the TCR result (1.8 K) is the same as IPCC. Both are consistent with the IPCC uncertainty ranges.

6/ The author argues that his estimate of ECS is "very consistent with the AR6 estimate". Yet his estimate of the aerosol cooling (-0.19°C ± 0.14 °C) isn't consistent with the IPCC estimate (see bullet A.1.3 of the AR6 SPM that says "The likely range of total human-caused global surface temperature increase from 1850–1900 to 2010–2019 is 0.8°C to 1.3°C, with a best estimate of 1.07°C. It is likely that well-mixed GHGs contributed a warming of 1.0°C to 2.0°C, other human drivers (principally aerosols) contributed a cooling of 0.0°C to 0.8°C, natural drivers changed global surface temperature by –0.1°C to +0.1°C, and internal variability changed it by –0.2°C to +0.2°C"). Thus IPCC estimate is centered on -0.4°C rather than -0.2°C. It would be good to mention and discuss this apparent disagreement, otherwise there is little value in flagging the agreement on the ECS.

The adjusted NTCF-induced cooling is now just within the very wide 5-95% uncertainty range of the IPCC estimate (adjusted for different reference periods and accounting for ozone, 0.15-0.57 K) but indeed skews towards a smaller absolute best-estimate value than IPCC. I cannot fully explain this but note that excluding models that do not satisfy additivity considerably improves agreement amongst the remaining 7 models regarding the best-estimate cooling due to aerosols.

The smaller best-estimate aerosol-induced cooling suggested by the regression analysis is not entirely surprising. For individual models this has been noted, also for the group (by Gillett et al., 2021) but not quantified in quite the same way as I do here. Quite what physical shortcoming is causing this, I can only speculate.

7/ Bottom-up estimates of the aerosol radiative forcing (e.g. Bellouin et al., doi: 10.1029/2019RG000660, 2019) are also larger (i.e. more negative) than implied by the aerosol cooling inferred from this study. This does not invalidate the current analysis. Yet the fact that it is at odds with a number of other studies requires some discussion.

I now discuss Bellouin et al. They come up with upper and lower bounds of the aerosol-induced radiative forcing of -2 to -0.35 W m^{-2}, translating into roughly 0.18 to 1 K of cooling due to aerosols between 1850 and 2005-2015. Thus the reference periods are very similar to the one used here (1850-1899 to 2000-2014). Their PDF for the combined aerosol-radiation and aerosol-cloud effect peaks at about -0.7 W m^{-2} or 0.35 K of cooling. My result is consistent with but skewing slightly smaller than Bellouin et al. I note that in the calculation, using only three models, where ozone is

explicitly included, I obtain 0.27+-0.08 K as the best estimate for the aerosol-induced cooling, close to the centre of the PDF of aerosol-induced cooling presented by Bellouin et al.

8/ Ocean diffusivity also affects the time evolution of the GSAT in models. How would biases in ocean diffusivity in the models affect the author's analysis?

I cannot say anything definitive about the role of ocean diffusivity. Models used here generally parameterize ocean eddies because of insufficient resolution. This may be a reason for general shortcomings in the models, e.g. misrepresentations of meridional heat flux in the ocean and consequent biases e.g. of high-latitude SST. However such process-oriented questions are far outside the scope of this paper and would require a process-oriented study to address.

9/ Lines 70-71, equations 4 and 5: the author appears to assume that alpha1 and alpha2 (beta1 and beta2) are independent variables so that their error variances can be summed but is this really the case?

These equations are now gone due to a change in method. Hence this comment is now moot. (I suppose they are actually not independent, so this wasn't quite the correct approach.)

10/ The reader needs to understand how DAMIP ensemble members for a given model were treated in the analysis.

The analysis only uses ensemble means. So I have formed, individually for every model, an ensemble mean of the available ensemble members for ts and tas (numbers are in table 1), and then merged these fields into the temperature field used in the rest of the analysis. This is explained in the text (around eq. 1).

Minor comments

Lines 19-20 is a bit of a truism.

Maybe but it remains correct. I have added that cumulative model improvements have not changed this situation.

Line 28: the decreasing aerosol loading is not a "feedback" but a forcing.

I have replaced "feedback" with "effect", i.e. how much warming this will produce.

Lines 33-34: sentence unclear, please reformulate.

I have reformulated the sentence. Hopefully it is now clear.

Figs 1 and 5: Plotting b versus a generally implies that b is on the y-axis. Doing the opposite is confusing.

I have reversed the axes.

Line 116: why four models and not all models?

This plot is now showing only 1 model (HadGEM3) to explain the concept. All other models are shown in figure A1. The aim was not to overload the figure.

Fig 2: is the plot for the model ensemble mean or for a particular member of DAMIP?

The plot is for the ensemble mean, as stated in the caption.

References

I am now discussing most of the below references, particularly Bellouin et al., Gillett et al., and Storelvmo et al..

Bellouin, N., J. Quaas, E. Gryspeerdt, S. Kinne, P. Stier, D. Watson-Parris, O. Boucher, K.S. Carslaw, M. Christensen, A.-L. Daniau, J.-L. Dufresne, G. Feingold, S. Fiedler, P. Forster, A. Gettelman, J. M. Haywood, F. Malavelle, U. Lohmann, T. Mauritsen, D.T. McCoy, G. Myhre, J. Mülmenstädt, D. Neubauer, A. Possner, M. Rugenstein, Y. Sato, M. Schulz, S. E. Schwartz, O. Sourdeval, T. Storelvmo, V. Toll, D. Winker, and B. Stevens, Bounding aerosol radiative forcing of climate change, Reviews of Geophysics, 58, e2019RG000660, doi: 10.1029/2019RG000660, 2019.

Charles, E., B. Meyssignac, and A. Ribes, 2020: Observational Constraint on Greenhouse Gas and Aerosol Contributions to Global Ocean Heat Content Changes. J. Climate, 33, 10579–10591, https://doi.org/10.1175/JCLI-D-19-0091.1.

Gillett, N.P., Kirchmeier-Young, M., Ribes, A. et al. Constraining human contributions to observed warming since the pre-industrial period. Nat. Clim. Chang. 11, 207–212 (2021). https://doi.org/10.1038/s41558-020-00965-9

Knutti, R. (2008), Why are climate models reproducing the observed global surface warming so well? Geophys. Res. Lett., 35, L18704, doi:10.1029/2008GL034932.

Storelvmo, T., Leirvik, T., Lohmann, U. et al. Disentangling greenhouse warming and aerosol cooling to reveal Earth's climate sensitivity. Nature Geosci 9, 286–289 (2016). https://doi.org/10.1038/ngeo2670