Christopher Smith: (The reviewer's comments are in red, responses in black.)

I thank Christopher Smith for some expert, balanced, constructive comments.

In this paper, Morgenstern uses 15 CMIP6 models contributing to the Detection and Attribution Model Intercomparison Project (DAMIP) to estimate the relative contributions to greenhouse gas and aerosol forcing in the present day, and uses the analysis to provide an emergent constraint on the equilibrium climate sensitivity (ECS). The two main findings of the paper are that (1) the ECS is in line with the IPCC headline assessment (likely range 2.5-4.0°C), a little lower than the models' unadjusted values, and (2) the aerosol warming contribution is only one-third as large as the models' unadjusted values and substantially less negative than the IPCC headline assessment of -0.5°C (-1.0°C to -0.2°C very likely range). If the second point is correct, the implications for future climate change are huge, in the sense that the masked warming by aerosols is small and climate would not be expected to warm by a large amount in any future emissions mitigation scenario.

The analysis hinges solely on CMIP6 models. Therefore, there is a risk that the conclusions are over-confident.

That is indeed a risk. I have now modified the discussion to make explicit that there are different possible explanations for the findings, and also provide a more in-depth discussion versus the IPCC assessment and other lines of evidence. This shrinks the discrepancy versus the IPCC estimates. I have also added as a very well-performing model MPI-ESM1-2-LR which simulates an aerosol-induced cooling of just 0.3 K, and requires relatively small adjustments to bring its historical temperature evolution into agreement with HadCRUT5.

One weakness of a purely CMIP6 historical approach is that the models are all forced with the same (largely uncertain) spatial and temporal aerosol emissions data set. It's possible that the era of strong aerosol cooling in many CMIP6 models (the whole 20th Century in some models, but we see that many see a bit of a step change around 1950) could be an artefact of the forcing dataset. It could also be because the models are producing overly sensitive responses to aerosols during these time periods, which may also be a factor of some models having high climate sensitivity. I don't believe anybody has put forward a convincing argument one way or the other yet, though Smith & Forster (2021) and Flynn et al. (2023) have both tried to answer this question.

I now discuss this situation and the references in more depth. I don't have a conclusive answer either as to whether this is a general model shortcoming or a property of the CMIP6 emissions dataset. The dominance of models requiring scaling-down of aerosol-induced cooling may indicate the former.

The point I'm trying to make is that if the "shape" of the aerosol cooling time series differs between the models and reality, then an optimal fingerprinting approach may try to mitigate the effect by selecting regression coefficients β_1 that are less than one. This will reduce any error in the total warming time series when the individual components are summed up, and also implies that $\alpha_1 < 1$ to balance out the positive effect of the GHG warming. We indeed observe that $\beta_1 < 1$ in all models and $\alpha_1 < 1$ in all but two models. The author "normalizes" this approach by also determining the regression coefficients compared to each model's historical run, which is a good idea. However, interestingly again, the regression coefficients α_2 and β_2 are also usually less than one (described in lines 150-151 as a lack of additivity). This could be suggestive of the regression approach attempting to minimize residual errors caused by natural variability rather than a genuine lack of additivity, though it should be noted that the omission of ozone and land use forcings may not be insignificant.

To investigate this, perhaps a rolling mean filter applied to the T_{h*} terms in eqs. (1) and (2) could be investigated. CanESM5 hints at this effect: at 25 ensemble members, its model-derived internal variability is small, and it is the only model where α_2 and β_2 are both greater than (and are also quite close to) 1, and the noted HadGEM3's approximate linear behavior has a 60-member hist-aer ensemble to draw upon.

Small changes to the regression are now improving the situation. Also I now use the "historical" simulations to restrict the analysis to additive models. CanESM5 is never considered "additive". Towards the end of the period its simulated warming is larger than can be explained by aerosols and GHGs. I now discuss the role of ozone in some more detail. Land use changes are very small climate forcers, on the global scale (IPCC AR6).

The shape of the historical aerosol cooling is something that we investigated in Smith et al. (2021). If we allow this to vary more (taking CMIP6 as an ensemble of opportunity, fitting a non-linear functional form and sampling the parameters) then we can construct aerosol forcing histories that do permit strong cooling and are still consistent with observations.

Therefore, it is my working hypothesis that the author finds a weak contribution to historical aerosol cooling because the historical shape of aerosol cooling (and forcing) is a poor fit to that implied by global temperatures and not easily resolved using a linear combination of GHG and aerosol attributed warming, and not necessarily because the present-day level of cooling is incorrect in the models (though historically, it likely was in some).

This would undermine most attribution approaches that rely on linear regression. I note that now eight models satisfy additivity for both the GHG induced warming and the aerosol-induced cooling. For all these models the regression produces very close fits to the observed temperature.

I'm also curious about the slightly different estimate for the present-day aerosol cooling to Gillett et al. (2021), who also did an optimal fingerprinting approach with CMIP6 DAMIP models and found aerosol cooling to be -0.7 to -0.1 °C. In their fig. 2b, it can be seen that aerosol regression coefficients > 1 for some models, though typically they also are in the 0 to 1 range. It would be useful to compare the differences and methods between the two papers.

Gillett et al. include in their analysis all models without regard to additivity, but only with one ensemble member each. Their regression model is very similar to mine. Their figure 2b indicates pretty small regression values for some of the same models that feature in my analysis (ACCESS-ESM1-5, CESM2, GISS-E2-1-G, HadGEM3-GC31-LL) and values greater than 1 for MIROC6 that has smaller aerosol-induced cooling. So I think their analysis is consistent with me, factoring in that I use ensemble means which reduces single-model uncertainties.

I do not want to come over as overly critical. It is a thorough yet concise paper, mathematically rigorous but not over-complicated, and the figures, equations and structure are clear and logically organized. Given the sensitivity and importance of the topic, the results should be contextualized relative to the IPCC assessment, which used more lines of evidence than solely CMIP6 (analogous to the ECS).

Indeed. I now state several times how the results compare to the AR6 assessed values. I will leave it to AR7 to place these results into context with other lines of evidence (paleo climate, process-oriented studies, etc), a full discussion of which is out of bounds for this paper. I believe though that in studies such as Gillett et al. (2021), non-additivity means that several models have to be removed from the analysis, which can substantially affect the overall conclusions.

Minor comments:

Abstract line 2: suggest replacing “anomalously large” with simply “larger”. I don’t think ECS of 5.6 or 5.7 K can be categorically ruled out.

Indeed. IPCC calls simulations produced by these models “low-likelihood, high impact”. I follow your suggestion.

Line 35: “heuristic regression”. I might be showing my ignorance here but I don’t know what this is. It seems to be defined as a machine learning concept (https://dl.acm.org/doi/abs/10.1145/503810.503823). Was this the method used? Eqs. (1) and (2) look more like regular least-squares.

No, the naming coincidence is accidental. I now drop the word “heuristic”. Indeed the approach is an ordinary least-squares regression method.

Line 60: I wonder why 1920-2020 and not the whole time period. Are results sensitive to the start date? I imagine they’d be very different if you used 1970.

I have now changed to a “symmetrical” regression (by introducing intercepts) and have expanded the regression period to 1850-2020 (or 1850-2014, for four models). The results are qualitatively similar to using the later start date.

Line 72: “single variable uncertainties”: would this be standard error?

Indeed this is the standard error.

Line 219-220: This statement of models exceeding 0.5K cooling being unrealistic is too strong.

I have rephrased the line to make it more accommodating of other possibilities and avoid categorical statements like this.

Line 246: “global warming”: since we’re also talking about aerosol cooling, I suggest being more general: “anthropogenic climate change”.

The term “global warming” is used here deliberately because the paper is only concerned with surface temperature.

Sign convention for any time you talk about a cooling, e.g. lines 9, 218, 219: a minus cooling is a double negative.

Those “minus” signs are now generally removed.

References:

I now cite and use all of these references.

Flynn et al. (2023): https://acp.copernicus.org/articles/23/15121/2023/acp-23-15121-2023.html

Gillett et al. (2021): https://www.nature.com/articles/s41558-020-00965-9

Smith et al. (2020): https://acp.copernicus.org/articles/20/9591/2020/acp-20-9591-2020.html

Smith et al. (2021): https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2020JD033622