Reply to reviewer 1:

I thank the reviewer for this additional review and some very useful comments. Review comments are in red, my responses in black.

The author has modified his methodology following the reviewers' comments and as a result his study has become more rigorous and more convincing. There remains a couple of issues, and I recommend that the author addresses at least the first one before the manuscript be accepted.

First the author may have misunderstood my comment on the "amalgamation" of the surface temperature product. The problem is not so much using SST instead of SAT over the open ocean. After all, over the open ocean, the trend in SAT must be very close if not identical to the trend in SST. The problem is that some surface temperature products such as HadCRUT5 use SST over the open ocean but SAT over land and sea ice. As the fraction is sea ice has been decreasing over the years, this makes a small but documented difference. Hence Eq 1 on line 88 is not consistent with HadCRUT5 in that HadCRUT5 does not consider the land fraction but the land and sea ice fraction. This is described in the SI of the Morice et al (2021) JGR paper: "For the HadCRUT5 analysis the weighting is also dependent on sea ice coverage, with sea ice regions treated as land" and as result the f in their equation varies in time. The sentence about HadCRUT5 on line 85 is thus not correct.

I have corrected this issue. Now I perform the analysis according to Eq. 1 which is what the reviewer suggests. Along with an increase in ensemble sizes for some models due to a general trawl for new data that I have conducted for this revision, but also a decrease for some models with small ensemble sizes due to unavailability of the siconc/siconca variable, this change in the base data has caused some minor numerical changes of the results which are however not fundamental.

Second I am always a little dubious of the relevance of comparing a model ensemble mean of historical simulations (i.e. the average between many historical realizations) with the historical observations (i.e. only one possible realization), especially for models that show a large amount of low-frequency internal variability. Sampling (randomly) one member, as done in Gillett et al (2021) and in many emerging constraint studies, is probably no better. I am not sure what the best way of doing would be (some sort of a probabilistic approach?) but I think that this problem has plagued the community for too long. At least the author should recognize this issue.

I agree that this is an issue; it is now mentioned, but not comprehensively discussed, in the paper. A further development would be to turn the theory in this paper into a probabilistic approach as suggested. This is beyond this paper though. When I find time, I may write a sequel to this paper laying out this theory.

The historical aerosol cooling inferred by the author is on the low end of other estimates. This may well be true and this study is solid enough to be eventually published but I note that there continues to be conflicting evidence on the magnitude of the aerosol forcing (and its contribution to cooling) during the historical period. This recent study (doi: 10.1038/s43247-024-01324-8) attributes the recent increase in the Earth Energy Imbalance (EEI) to fading aerosols and it is hard to reconcile with a small aerosol effect.

Thanks for pointing out this paper. Results in this paper at least superficially appear inconsistent with the results presented here. I can only acknowledge this problem; further research is needed to resolve it.

Minor comments

Line 23: p.a. => per annum

This is now spelled out.

Line 90: “surface temperature” should have read “land surface temperature”

I use the whole “surface temperature” field (ts) and the whole surface-air field (tas) but the weighting is such that over land, away from the coast, I actually use “surface-air temperature” not “land surface temperature”. I think the present formulation, with the corrected Eq. 1, is now clear.