We would like to thank the first reviewer, Christopher Smith, and the anonymous second reviewer for a careful and thorough reading of this manuscript, and for the thoughtful comments and constructive suggestions provided. These help to improve the quality of this manuscript. We provide point-by-point responses to each comment from both reviewers below.

Response to Reviewer 1

Reviewer Point P 1 — *This paper investigates whether aerosol forcing is the reason why the CMIP6 ensemble is cooler than observations in the middle of the 20th century, as opposed to the CMIP5 ensemble that warms roughly in line with observations. This is a topic that continues to intrigue researchers and one that it is important to try to understand, in order to potentially correct model biases in the future. Unfortunately, the authors could not reach a definitive conclusion around what causes the mid-century cool period in CMIP6 models, finding it is not solely due to aerosol forcing which is the obvious candidate, but aerosol forcing is likely to be one factor of many. In this regard, they reach similar conclusions to Smith & Forster (2021). Despite a null result, this is a useful contribution to the literature and will hopefully motivate other researchers to continue to study the topic.*

Reviewer Point P 2 — *Comments are mostly minor, but please note for the ERFari values reported in this study, unfortunately the method I used in Smith et al. (2020) was slightly flawed! See Zelinka et al. (2023), https://egusphere.copernicus.org/preprints/2023/egusphere-2023-689/egusphere-2023-689.pdf which explains and corrects this. Updated values are in table 2 of that paper. I trust that using corrected values will not change the results of this paper substantially.*

Reply:

We thank the reviewer for the positive comments, and though the main conclusion of our manuscript is a null result, we agree that it presents a useful contribution to the literature. We also hope that our work will motivate other researchers to continue to investigate this problem. We also thank the reviewer for bringing to our attention that the ERFari values that we used have been corrected, and we have updated our manuscript accordingly.

As a consequence of using the corrected ERFari, the association that we originally obtained between the mid-century warming and ERFari magnitudes for Block 3, which was shown in the former Figure 5, has disappeared; the corrected ERFaci, ERFari-LW or -SW, and ERFaci-LW or -SW magnitudes also did not show an association with the mid-century warming for any warming block nor for the entire ensemble. The former Figure 5 showing the mid-century warming vs. ERFari magnitudes has therefore been removed from manuscript, as well as the column for ERFari values in Table 1, Lines 74-75 citing the source of the original, erroneous ERFari values we used, and Lines 283-289 in which ARI as a potentially promising research direction is described based on the formerly strong association between warming and ERFari for the Block 3 models. However, our analysis does not definitively rule out a role for ACI or ARI in causing the suppressed warming, but rather that any potential role they play is more nuanced and may be important for only some models and not others. In some ways, the much weakened association between the warming and ERFari for the Block 3 models makes the question of the causes of the suppressed warming more tantalizing for researchers as the causes seem to be quite subtle and not obvious.

Reviewer Point P 3 — *Line 4: "observed anomaly" – which period are we talking about here?*

Reply:

We have added the year range, 1940-1970, over which we computed the anomalies in parentheses to Line 4 for clarity.

Reviewer Point P 4 — *Line 13: "encouraging" – why? Either that there is some consistency that hints at a constraint, or that weak aerosol forcing is good in the sense that it implies a smaller committed warming (Watson-Parris & Smith 2022, https://www.nature.com/articles/s41558-022-01516-0)?*

Reply:

We intended "encouraging" originally in terms of a consistency that hints constraint, though it would be a weaker constraint than hoped, on the aerosol forcing. Taking a broader view, we agree with the reviewer that the tendency for more realistic mid-century temperature anomalies to be obtained for models with weaker aerosol forcing is also encouraging in that that implies a smaller committed warming.

Reviewer Point P 5 — *Line 15-17: Fully agree with this statement*

Reply:

We thank the reviewer.

Reviewer Point P 6 — *Line 31: Bellouin et al. (2020); also Forster et al. (2021), the AR6 WG1 Chapter 7 assessment, came to a very likely range for 1750 to 2005-14 of -2.0 to -0.6 W m-2, again from multiple lines of evidence; the Bellouin et al. paper put in much of the foundations for this work.*

Reply:

We have updated Line 31 to also include the range -2.0 to -0.6 W m-2 for the forcing.

Reviewer Point P 7 — *Line 38-39: Smith et al. (2021; https://agupubs.onlinelibrary.wiley.com/doi/full/10.102 also found the Stevens model to be overly simplistic and could not capture the diversity of historical aerosol forcing, so proposed two modifications: the addition of additional species, and relaxing the constraint that the aerosol indirect effect depends on natural emissions and replacing this with a generalised shape factor, allowing forcing to scale logarithmically (as proposed by Stevens) or approximately linearly (as proposed by Booth and Kretzschmar) with emissions, depending on parameter choices. With this model it was easily possible to obtain stronger aerosol forcing than -1.0 W m-2 that was consistent with historical warming.*

Reply:

We thank the reviewer for bringing Smith et al. (2021) to our attention, and have included this information and a reference to this paper following Lines 38-39 in our manuscript

Reviewer Point P 8 — *Line 74: following comment above, best to take results from Zelinka et al. (2023).*

Reply:

Please see our response to Points 1 and 2 above.

Reviewer Point P 9 — *Line 91: is it necessary to exclude 1963-66? Since the simulated climate projections from CMIP6 should have included volcanic forcing too and hence the contribution from Agung is present in both the observations and the models.*

Reply:

We think it is necessary to exclude these years, and have done so in the computation of the anomalies from both models and observations, in order to be sure to remove volcanic influence and isolate anthropogenic aerosol forcing. This was also done for consistency with Flynn and Mauritsen (2020) where we identified the suppressed warming, in which we excluded years of significant volcanic eruptions from our anomaly calculations and historical time series.

Reviewer Point P 10 — *Line 115: Would it be better to use piClim-anthro? The sum of piClim-ghg and piClim-aer excludes contributions from land use change and ozone. It also excludes natural forcings, though there is not a time slice in RFMIP available to estimate it and it's fair to assume there wasn't a big change from 1850 to 2014.*

Reply:

It is certainly very useful to examine the piClim-anthro simulations, as well as the RFMIP simulations designed to isolate forcing due to forcing agents other than those examined in our manuscript; we encourage researchers to take advantage of the rich suite of model experiments performed as part of RFMIP. However, in this manuscript, our primary focus was on anthropogenic aerosol forcing due to its large uncertainty and important influence on the 20th century warming evolution, and secondarily on the forcing from the well-mixed greenhouse gases, as these are of course the main drivers of the warming and differences in well-mixed greenhouse gas forcing among CMIP6 models could play in a role in the suppressed warming. This should not be taken to mean that other forcings, such as from land use change or the short-lived greenhouse gases, are not important and could not have influenced the warming evolution.

Reviewer Point P 11 — *Line 119: I'm not sure I understand the drift correction method in the piClim-ghg and piClim-aer experiments. The piClim-control is only 30 years for most models so I'm not sure there are many branch points for piClim-ghg and piClim-aer. As the ERF calculation uses fixed SSTs this should also remove the need for drift correction. For some forcings there is a relaxation time where the atmospheric response to a forcing is not instant; fig. 2 of Smith et al. (2020) shows this in action (CNRM-ESM2-1 aside).*

Reply:

We thank both reviewers for catching this; inclusion of how we handled drift in the RFMIP piClim-control simulations is a result of copy and paste from an earlier draft and then not careful enough editing of this section by the main author. We did subtract the piClim-control simulation as-is rather than performing a drift correction, as is necessary for other types of CMIP simulations, and so have deleted the lines about drift correction from Section 2.2.

Reviewer Point P 12 — *Line 184: Clear sky flux change is not the same as ARI. However, they are highly correlated, so I suppose you can use clear sky flux change as a proxy for ARI. Section 4.3*

in Zelinka et al. (2014) gives a good discussion (https://agupubs.onlinelibrary.wiley.com/doi/full/10.1002/2014JD0
I did calculate ERFari from 13 CMIP6 models (using the correct version of the APRP code) here, if you want to use it: https://github.com/chrisroadmap/cmip6-aerosol-forcing/tree/main/output

Reply:

We agree that clear-sky flux changes and ARI are not the same, though it is true that they are highly correlated, as ARI is masked by clouds and should therefore be smaller than the clear-sky change. However, we plot the clear-sky fluxes from the piClim-histaer experiments specifically in this figure, so only aerosols could impact flux changes in this figure and Line 184 and that is why we highlight ARI as an example.

Reviewer Point P 13 — *Line 204: this is true in the ensemble of opportunity that CMIP6 models provide: a sample size of 36 models, only about half of which can give you an estimate of present-day aerosol forcing, even fewer give you an estimate of the aerosol forcing during the period of interest. Although not stated, I'm uncomfortable in claiming this to be a true result in the real world, as we showed in Smith et al. (2021).*

Reply:

We agree that our ensemble size is too small to claim this result to be true in the real world, and this is one reason we hope that more models will perform and share RFMIP-type experiments for researchers to analyze. We perhaps were not clear enough in Line 204 that we are speaking certainly only about our model ensemble, and not all models, and have modified Lines 203-204 to say "...no model within this subset with strong aerosol forcing..."

Reviewer Point P 14 — *Line 243: no fault of the authors, but some of the results in the paper suffer from a lack of participating models in each Block, showing again how important that models run the ERF experiments from RFMIP.*

Reply:

We wholeheartedly agree, and hope that our work can contribute to motivating other researchers to perform these experiments and researchers to analyze them. Our results point to some interesting conclusions, but more models performing RFMIP-type experiments are needed to increase confidence in our results.

Reviewer Point P 15 — *Line 252: this suggests that* ***a low*** *greenhouse gas forcing...*

Reply:

We agree, and have modified Line 252 to reflect that it suggests a weak greenhouse gas forcing.

Reviewer Point P 16 — *Line 282: it remains a mystery. Would the pattern effect have anything to do with it? I'm not sure how this study would evaluate this. Smith et al. (2021) included the effect of a forced pattern effect from the increasing climate sensitivity over time as simulated by an ensemble of energy balance model simulations trained on CMIP6 models, but did not evaluate this effect either. I could see how a strong aerosol forcing could be consistent with a virtually non-existent historical pattern effect, or a weak aerosol forcing masking a strong pattern effect. AMIP experiments (Andrews et al. 2018, https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2018GL078887)*

point towards a strong historical pattern effect, maybe adding weight to the suggestion that aerosol forcing may be on the weak side.

Reply:

The pattern effect could also be at play in causing or modifying the suppressed warming, and our study cannot say either way if or how it would be involved. We performed a rudimentary analysis of the relationship between the mid-century temperature anomalies and a metric representing the pattern effect based on work performed by one of our co-authors on this manuscript, but the results were inconclusive, and so we did not go farther with analyzing the pattern effect for this study. We hope that researchers will investigate any potential links between the suppressed warming in CMIP6 models and the pattern effect.

Reviewer Point P 17 — *Line 283: again unfortunately the ARI is wrong in Smith et al. (2020).*

Reply:

Please see our response to Points 1 and 2 above.

Reviewer Point P 18 — *Line 291: this should not be that surprising. Energy budget arguments can permit a present-day aerosol forcing as strong as -2 W m-2, as discussed in Bellouin et al. (2020) and AR6; Smith et al. get quite close with a 5th percentile of -1.8 W m-2. The time evolution of the historical forcing matters, not just its present day value, and pattern effect probably matters too.*

Reply:

We accept the reviewer's point here; we were mainly thinking of the argument that aerosol forcing stronger than -1.0 W m-2 should be unable to produce realistic warming, rather than the Bellouin et al. (2020) or AR6 ranges. We have modified the beginning of the Conclusions paragraph starting with Line 291 to reflect that it is not surprising that some models with strong aerosol forcing were able to produce mid-century warming (though often unrealistic warming), and that the more surprising result was the large degree of suppressed warming or even mid-century cooling obtained for some models with weak aerosol forcing.

Response to Reviewer 2

Reviewer Point P 1 — *This study presents an analysis of the mid-century (1940-1970) "suppressed warming" or cold temperature bias in the CMIP6 generation of models versus CMIP5 and observed warming. The authors partition the CMIP6 models into 3 blocks depending on whether the simulated mid-century temperature anomaly is above the CMIP5 mean; is between the CMIP5 and CMIP6 means or is below the CMIP6 mean anomaly. The relationship between these blocks of models and the effective radiative forcing from both aerosols and well-mixed greenhouse gases is examined. While block 1 models have the most realistic mid-century temperature anomaly and weakest aerosol forcing, there is no meaningful difference between the other 2 blocks. Block 3 in particular deviates from the expected temperature anomaly-aerosol ERF relationship simulated by sensitivity runs conducted with the MPI-ESM1.2-CR model. This points to other causes, outside of*

the aerosol ERF, of the cold biases in these models. In addition no meaningful correlation is found with the WMGHG ERF. While not the golden nugget in understanding the underlying cause(s) of the anomalous historical cooling that is a trait of many CMIP6 models, this is a useful addition to the expanding literature discussion on this topic. It is a topical, and relevant study and would be suitable for publication in ACP following consideration of minor comments below.

Reviewer Point P 2 — *My main over-arching concern regarding the main Conclusions of this paper is around the limited number of models in the CMIP6 database that have carried out these, one could argue, essential simulations. This number is further reduced if you break it down into models which interactively simulate the evolution of aerosol versus those that simply prescribe aerosol concentrations. While the authors acknowledge this in their results it very much limits the significance of the findings, in particular wrt Block 1 which only includes 3 models. The reader should be very clear on these limitations, it should be clearly mentioned/repeated in the Conclusions and the authors should highlight the requirement for more models to do these experiments.*

Reply:

We thank for the reviewer for the positive comments. We entirely agree that the small sample sizes of models performing these RFMIP experiments place limits on the confidence in our results, and we hope to see more models performing such experiments for the community to analyze. We are very curious ourselves to see if the findings we present in this manuscript hold when more models are included. The reader should absolutely keep in mind the small sample sizes, and we have tried to emphasize that throughout the manuscript. We have added additional clarity on this to the reader in the Conclusions after Line 273, as well as the recommendation suggested by the reviewer for more models to do these experiments.

Reviewer Point P 3 — *L80 : What is the limitation of only using the first ensemble member of each models historical ensemble? This might not necessarily represent the ensemble mean temperature anomaly and so could impact the results. Have the authors examined this?*

Reply:

This was done to maintain consistency with Flynn and Mauritsen (2020), where we first identified the suppressed warming in the CMIP6 ensemble-mean. However, we did some investigation of the impact of examining more than just the first ensemble member for a model, and this was not found to have much impact on the temperature anomaly for most models.

Reviewer Point P 4 — *L85 : There is a more recent version of HadCRUT data (HadCRUT5) that is better to use (https://doi.org/10.1029/2019JD032361)*

Reply:

We thank the reviewer for bringing this to our attention. However, while we used the Cowtan and Way Version 2.0 observationally-based in-filled global temperature data set in our manuscript that is based in part on HadCRUT 4.2.0, we think it is unlikely that updating the observationally-based data set will significantly alter our findings in the analysis we performed. This is because the observational warming anomalies are not a large focus of our manuscript and are not used to divide the CMIP6 models into warming blocks, but rather to say which models in our subset were more or less realistic. Using the more recent data set is unlikely to change that, given that most of the unrealistic models were quite unrealistic in either direction (too warm or too cold during the mid-century period).

Reviewer Point P 5 — *L116: I don't understand why the models exhibit drift, these should be 30 year fixed SST timeslices? There should also be no branching necessarily (L119 in the perturbed forcing experiments as they are just parallel experiments?*

Reply:

Please see our response to Reviewer 1, Point 11, above, as that reviewer raised the same concern.

Reviewer Point P 6 — *Figure 3: It would be informative to include the Block 1/2/3 mean values here as well as the observed temperature anomaly for reference.*

Reply:

We have added the mean temperature anomalies for Blocks 1, 2, and 3 as horizontal lines just as for the observations to Figure 3.

Reviewer Point P 7 — *L193: misspelt aerosol*

Reply:

We have fixed the misspelling.

Reviewer Point P 8 — *L202-204: I don't understand what is meant by this sentence, consider rephrasing.*

Reply:

We have modified this sentence to make our meaning more clear to the reader.

Reviewer Point P 9 — *L214: I think one of the findings / hypotheses of the Zhang paper was the role of process complexity (ie: earth system additional process such as fully interactive chemistry influencing aerosol composition and subsequent forcing versus physical climate models with lower complexity) and not just aerosol emissions.*

Reply:

We agree with the reviewer that the role of aerosol emissions was not the only important finding of Zhang et al. (2021) and did not intend to imply so; we just picked one of the findings of important differences among models identified by this paper, as it leads nicely into Section 4.2. However, we have included the use of fully interactive chemistry in ESMs compared to GCMs as an additional example of important model differences identified by the Zhang paper to Lines 214-5 to avoid the implication that only the type of aerosol input used matters for the suppressed warming.

Reviewer Point P 10 — *Section 4.2: I find this section a little limited/simplistic as there could be many other sources of systematic differences in these models outside of whether it takes aerosol emissions as input or not. For instance, the level of complexity of the aerosol scheme, how the ACI and ARI are represented not to mention the role of other parts of the physical system which could influence surface temperatures, eg the role of the ocean. These should at the very least be mentioned.*

Reply:

We agree with this comment, though investigating such differences among models if it is not readily available in an accessible, easy-to-decipher format, such as the model configuration documentation

provided as part of CMIP, can be quite tricky and time-consuming. However, our analysis should not be taken to mean that other differences among models, such as complexity of the aerosol scheme and how ACI/ARI are represented, cannot be not involved in causing the suppressed warming. Indeed, we encourage researchers to investigate other causes, and for modeling groups to make such information easily available, when possible. We have added a mention of these other causes to the end of this section following the reviewer's suggestion, and clarified that we cannot comment on their involvement in the suppressed warming.

Reviewer Point P 11 — *L222: should read: a standardized anthropogenic aerosol emissions input data set. Not all natural emissions were standardized across the models.*

Reply:

We thank the reviewer for catching this, and have updated Line 222 accordingly.

Reviewer Point P 12 — *L223-230 It might be better to visualise the prescribed concentration vs emissions-based models. Figure 3 could be repeated for instance changing the colours of data points to represent PC or E or PC/ice or E/ice, where ice = aerosol-ice interaction.*

Reply:

We have taken the reviewer's suggestion and repeated the former Figure 3 but with the marker shapes representing PC or E rather than aerosol-ice interactions or not, and included a reference to this new figure within Lines 223-230. This new figure is now Figure 3, and the original Figure 3 is now Figure 4, to follow the order of discussion of aerosol input data set type and inclusion of aerosol-ice cloud interactions within this section of the manuscript.

Reviewer Point P 13 — *Why is Figure 5 only mentioned in the Conclusions? Should this not be part of Section 4.1?*

Reply:

We did initially place Figure 5 in Section 4, but given that our main conclusion is a null result, we wanted to end the manuscript on a more encouraging note by presenting one potentially promising direction for further investigation of the causes of the suppressed warming. Thus, we moved Figure 5 to the Conclusions for that purpose.

However, please see our response to Points 1 and 2 from Reviewer 1 above; it turns out that the ERFari values we used were calculated incorrectly, and use of the corrected values greatly weakened the association between ERFari and the mid-century warming for the Block 3 models, and thus removed ARI as an obvious potentially research direction. This further points to a more subtle and nuanced cause or set of causes for the suppressed warming.

Reviewer Point P 14 — *Finally, consider citing the paper of Mulcahy et al. (2023) (https://doi.org/10.5194/gm 16-1569-2023). This updated configuration of UKESM1 shows a significant improvement in the cold historical temperature bias despite only a small change in the present-day aerosol ERF. This directly supports this study and points to a more nuanced process driving surface temperature response in the fully coupled system.*

Reply:

We thank the reviewer for bringing this paper to our attention, and have now included it in the Introduction after Lines 55-58, as we agree that it does support our findings indicating a more nuanced cause(s) driving the suppressed warming.