We thank the editor and the reviewer for their careful reading. We address each of the points raised below. Our comments are in blue below, first to the editor's comments and then to the reviewer's.

I apologise for the delay in returning this manuscript to you. One of the two reviewers of the revised submission still raises concerns on the lack of contextualisation of the uncertainties/limits of the significance testing in your results. In your replies, you argued for the fact that large enough differences are robust even between individual model ensemble members. This may be true, but I don't find the authors' replies on this point to be entirely satisfactory. If one takes a single CMIP6 model run, for example, some measure of variability around the mean would likely result in a narrower range than the uncertainty deriving from running 30 or more model ensemble members, thus giving statistical overconfidence in your results. Another example of perhaps overconfident statistical testing, is the use of Welch's t-test. The authors never specify whether this assumes that the data being tested follows a specific underlying distribution, and never verify whether their data satisfies this. They also quote Wilks, but never explicitly mention whether a multiple testing correction is implemented (as per the Wilks 2016 "Stippling" paper).

I do not agree with the Reviewer that the above is sufficient grounds for rejection, but it is definitely something that should be addressed with a combination of careful rewording - to acknowledge that there are, like in all studies, some caveats to the conclusions being drawn - and reconsideration or further justification of some of the adopted statistical testing approaches. I also ask the authors to provide point-by-point replies to the rest of the remaining Reviewer comments.

We agree with the editor on the importance of including appropriate caveats on use of statistics and have added several clarifying comments and cautions. The reviewer's primary concerns are around model error in relying on a single climate model, and we have added further emphasis on this caveat in the abstract and conclusions in particular.

Regarding the editor's concerns on statistics:

(1) First, regarding the number of ensemble members, it is certainly true that when one has very few data points one does not obtain an adequate estimate of the standard deviation, leading to overconfidence (or sometimes underconfidence). Here, we have 3 ensemble members and look at 20 years of simulation output for each variable, so we have a total of 60 data points. Accounting for autocorrelation in the time series results in fewer independent data points, with the exact number dependent on the specific variable being considered; we believe that this is sufficient to estimate the standard deviation (which is then used by the t-test to assess whether differences are statistically significant). Note that while we disagree with the reviewer on the ability to define a single universal number of data points for all

statistical testing in climate science (simply because that number clearly depends on the signal, so the size of the forcing and how that forcing affects a particular variable, and also on the variability of the variable in question), the number of independent data points here is indeed larger than the estimate from Bittner et al that the reviewer suggests.

(2) Second, a t-test assumes that normality is a reasonable approximation for the underlying probability distribution; this is a good assumption for annual-average outputs from climate models (due to the central limit theorem). We add a note to this effect.

(3) Third, we did not use a multiple testing correction but we have added a caution in several places regarding the fact that we test many different variables and that as a result it is possible that some differences may appear to be statistically significant simply by chance. In particular, at the beginning of the results section where we describe the statistical testing used herein, we add "We use t-tests to estimate significance, which assume that variability is approximately normal; this is a reasonable approximation for annual-mean climate variables. One caution on interpreting results is that in evaluating many different climate variables, some will appear to be statistically significant at the 95% level by random chance (Wilks, 2016)." We add a similar comment at the end of the Summary section.

Comments on "Hemispherically-Symmetric Strategies for Stratospheric Aerosol Injection " by Zhang al.

Zhang et al document the response of the climate system for a given stratospheric aerosol injection (SAI) strategy. To maintain the same mean surface temperature, the authors design several SAI strategies to be studied using the CESM2 Earth System Model (WACCM6-MA) and compare the results with the reference scenario, the SSP2-4.5 global warming scenario. My main concern with this paper is the lack of caution in the way the results are presented here as well as the lack of context on the complexity of our climate system and associated feedbacks, stratospheric circulation and variability as well as model biases in terms of precipitation and the lack of model representation of our complex system. I highlighted this point in my previous review, but it is not seriously addressed. The review does not address all my concerns, which is why I recommend major revisions. From a single model study, the authors intend to generalize their experiments to all models. The conclusion and abstract still marginalize the impact of SAI on precipitation and it even seems to me that the authors exaggerate their results by saying that SAI is good because it decreases temperature while ignoring the importance of the impact of SAI on precipitation and the related implication on food security, agriculture and so many others vital component for human survival. I am therefore going to reject this article and encourage for resubmission after addressing these serious issues.

While we did include caution regarding the broader interpretation from a single-model study, we have made this caveat more explicit both in the abstract and in the introduction, where we add a sentence about using these results to motivate model intercomparisons. The final paragraph of the conclusion section already emphasizes that the results are obtained in a single model and that multi-model explorations are critical; the penultimate conclusion paragraph further articulates that one motivation for the current study is to support GeoMIP model intercomparisons. We thus feel that this point is already well emphasized in the conclusions, though we add a further note about the need for multi-model studies in the first paragraph of the conclusions as well.

Note, of course, that the same concern regarding use of a single model could be made regarding the majority of papers in climate science. At no point do we claim that all models would behave the same way as this one, and in fact one of the points of our study is to introduce a common framework to use in future model comparisons in which inter-model biases and differences can be assessed explicitly (noted in the conclusions). In addition, one of the motivations for considering only hemispherically-symmetric injection strategies is that the asymmetry required to compensate for interhemispheric asymmetry in temperature is expected to be model-dependent, as we note on line 105 of the previous version.

Further, the paper never says that SAI is “good” (or similar normative statements) but simply objectively reports findings on temperature, precipitation and other climate metrics. Further analysis of the associated changes in stratospheric and tropospheric temperatures, circulation and chemistry, and their links to the surface climate responses in these simulations is made in our other recently published study (Bednarz et al., 2023), to which we add a citation at the end of the introduction. We also note that in response to the reviewer’s previous review comments, we greatly expanded the discussion of regional precipitation responses including additional figures on the differences in response over the Congo and Amazon basins; as noted previously we thank the reviewer for pointing out that insufficient emphasis in our original submission.

Major points:

1. The authors did not address this major issue: “The surface climate response to different SAI strategies is present without a clear understanding of the impact of model internal and inter-annual variability on the distribution of SAI in the stratosphere as well as its feedback on the surface climate. According to Bittner et al. (2016), 7 ensembles in the tropics and 40 ensembles in the extra-tropics are needed to accurately capture the model circulation response to SAI, and hence the corresponding feedback on surface climate. Caution should be exercised in discussing the results here. Three ensembles are not enough to constrain the internal variability of the model”. I invite them to read the Bittner et al 2016.

We disagree with the appropriateness of choosing a single universal number for the “correct” number of ensembles and the relevance of the specific quantification from Bittner et al to our work. Bittner et al. (2016) shows that a large number of ensemble members are required to detect NH polar vortex strengthening in the first post-Pinatubo winter. This number depends on

the latitude considered and ranges from 7 at the southward flank of the maximum positive wind anomaly to more than 40 members at high latitudes." (Section 4 of Bittner et al.).

First, our current study does not deal with the changes in the NH polar vortex, the variability of which can indeed be particularly large (we note that this aspect of the climate response to SAI in these simulations has been assessed in our other paper, Bednarz et al., 2023, which has now undergone a thorough peer review and is published).

Second, our study does not analyze changes in a single year or a single season as done in Bittner et al. study - in fact we analyze changes over the means over the 20-year long period and 3 ensemble members, and so we have effectively 60 semi-independent data points, which is indeed larger than the number the reviewer argues is necessary. (The effective degrees of freedom need to be adjusted to account for autocorrelation in the time series, but this still leads to more than the 40 independent degrees of freedom noted by the reviewer for every variable except global mean temperature.)

Third, the ability to distinguish a response depends on the magnitude and duration of the perturbation, and as such there are large differences between a short-term volcanic forcing and a continuous SAI.

Lastly, the reviewer is correct in that more ensemble members gives better ability to distinguish signal from variability. However, there is no single threshold that can be defined for the required number of ensemble members in general, and thus generalizing the single model and single-forcing study by Bittner et al. to all models and all forcing scenarios is not appropriate.

The effects of internal variability are already included in all analyses presented, all of which include error bars indicating the limit of confidence resulting from variability, as is standard practice. We account for autocorrelation in the time series. We have also added a comment in discussing statistical testing in general about the assumption of normality embedded in using a t-test (which is a very good assumption, following from the central limit theorem), and about the caution in evaluating statistical significance for many different variables that some may appear to be significant by chance; this latter comment we also reiterate at the end of the Summary section.

2. The abstract does reflect the content of the paper. Therefore, it needs to be rewritten

We respectfully disagree with the reviewer. If there are any specific aspects of the abstract that do not reflect the content of the paper, we would appreciate it if those could be pointed out.

3. Page 1, line 20, please added after "latitudes" this "based on a single model study".

To improve the emphasis in the abstract, we add "in one climate model" earlier in the abstract on line 7. On line 20 this is already abundantly clear from context as the sentence is describing

the results in the paper. (Nonetheless, the sentence noted here should reasonably be expected to hold in any climate model.)

4. Introduction, please add a section on model limitation and biases regarding SAI and precipitation before “Defferent SAI strategie…” Models in general even struggle to reproduce Pinatubo or later volcanoes. Current climate response to SAI are still model dependent due to unconstrained internal model variability as well as interannual and decadal variability of the climate system.

CESM(WACCM) does a reasonable job of matching stratospheric aerosol distribution after Pinatubo, as was investigated by Mills et al. (2017); we have now added a sentence noting this, in Section 2 where the climate model used herein is described. There are multiple examples over the past several decades of models actually reproducing quite well both Pinatubo and later volcanoes - both in terms of the aerosol distribution, and in terms of the overall climate response. Of course we agree that any model projection should be interpreted as the predictions of a model.

5. Page 1, line 1 please added after “scenario” this “based on a single model Study”.

The word “scenario” does not show up on page 1, line 1, and we are not clear what line the reviewer intended to refer to. We do add “in one climate model” on line 7 to be explicit earlier in the abstract.

6. Page 3, line 80 please add after “variability” tis “… and model biases”.

The suggested change would not be correct. The sentence refers to the ability to detect changes, which is limited by internal variability but not by model bias. Model bias is a source of uncertainty (and thus the differences found in one model may not be the same as what would happen in the real world). Errors in the model’s ability to represent natural variability accurately would also lead to an error in estimating detectability; factors like that are why the sentence already says “among other factors”.

7. Page 4, line 92, the linearity hypotheses is not taking into account the feedback processes (Stratospheric water vapor, O3 and SAI) on circulation and climate.

The sentence is true as written. The presence of feedback is irrelevant to whether linearity is adequate. What is relevant is whether the feedback is itself significantly nonlinear.
There are ample examples of feedbacks in linear systems (see, e.g., Astrom and Murray, https://fbswiki.org/wiki/index.php/Feedback_Systems:_An_Introduction_for_Scientists_and_Engineers or any other textbook on feedback systems). Ultimately the only question is whether or not linearity is an adequately good approximation as there is no physical system as a truly linear system. The adequacy of the assumption would need to be validated in this particular context, which is why it is noted in passing here simply as a possible choice of assumption in future work, but there is ample support for linearity being an adequate approximation for SAI in

particular (e.g., MacMartin et al. 2013, Irvine et al. 2019, MacMartin et al. 2019, Visioni et al. 2023, etc.) and so it is not unreasonable to expect that it will likely prove sufficient.

8. Page 4, line 120, “significant perturbation of the interhemispheric temperature gradient and the associated location of tropical precipitation” only correspond to fast response of the SAI as the aerosol can be transported into deep stratosphere then impact the opposite hemisphere few month later.

The reviewer’s claim is not correct. Injecting away from the tropics in a single hemisphere will put aerosols primarily in that particular hemisphere (constrained by Brewer-Dobson circulation), and indeed lead to an asymmetric aerosol burden and corresponding forcing, and a resulting shift in tropical precipitation. This is well documented in numerous papers including the one cited here (Haywood et al. 2013). We therefore retain the current wording.

9. Table1: How these aerosols and their life time are sensitive to the altitude of injection.

The reviewer is correct that the results will also be somewhat sensitive to the altitude of injection, although as long as the injection is not too close to the tropopause the main effect of altitude will be to change the injection rates required to achieve a given cooling and not the spatial pattern of that cooling (Lee et al. 2023); we now add a sentence describing the choice of altitude and the effect of that choice.

10. Page 9, line 223-224: This is due partly to the altitude of injection of SO2 as the shallow branch of the BDC will wash out the aerosols.

True that the altitude affects lifetime; the relevance of altitude is now noted earlier as commented on above.

11. Page 9, line 227, I am still not convince that the existing BDC asymmetry between SH & NH due to the wave activitives, Polar vortex, internannual variability modulation has no impact on the aerosol distribution in the lower stratosphere (Fu & Qu, 2013).

We are confused about this comment; of course BDC asymmetry affects aerosol distribution and we never claimed that it didn’t. But that is a small effect compared to the asymmetry arising from injecting 12.2 Tg/yr in one hemisphere and 5.8 Tg/yr in the other, which is what the line in question is pointing out.

12. Page 13, line 292, Please add after “climate.” this “based on the WACCM results”

The sentence in question is describing Figure 8; we believe that it is clear that Figure 8 is presenting results from the simulations described herein and that it is unnecessary to repeat “based on the WACCM results” in this and any other sentence in the paper that is describing simulation results. The reviewer’s suggestion would be appropriate in conclusions that could reasonably be inferred as making claims broader than these simulations.

13.Page 24, line 477 after “injection strategy” please add “ as well ad the models ability to capture the complexity of the Earth climate system and its variability, which is not investigated here.”

The sentence in question is accurate as written. That is, that we have to be able to assess the role of scenario and strategy as part of assessing possible outcomes of SAI. It is also true that we need good enough models, but that is not the focus of this paper - there are many papers on that subject already, while the purpose herein is to highlight the importance of the strategy dimension in particular. Much of the rest of the conclusion section already emphasizes the limitations of a single-model study and the need to conduct analyses in multiple models.

14.Page 24, line 477 please “based on a single model study” after “SAI strategies …”

We think the reviewer was intending to refer to line 478 (as “SAI strategies” does not appear in Line 477). The sentence in question simply notes that different SAI strategies will affect the surface climate differently; that is clearly a conclusion that would hold in any climate model as well as in the real world. Rather than adding a comment both here and in the next sentence as the reviewer suggests, we include an additional comment towards the end of the paragraph noting the need to conduct similar analyses in multiple models; this provides emphasis that the results are based on a single model. This point is now made in every paragraph of the Discussion section. We furthermore add this point in two places in the Summary section describing the results.

15. Page 24, line 481 please “based on a single model study” after “fundamental limits of SAI”.

We add an additional note about the importance of conducting similar analyses several sentences later “as well as to conduct similar analyses in other climate models”.

16. Page 25, line 487, what about precipitation results?

The previous paragraph notes that there are many climate goals that are relevant and could be optimized over, not just temperature or precipitation.