We thank the editor for their comments. Our response is in blue below.

1. Concerning the number of ensemble members, it would make for an interesting discussion point to mention the paper by Bittner et al. (2016) and explain – as you do in your replies – why those conclusions do not necessarily apply to your work. This can be very short, but may help attentive readers to better understand the rationale behind the robustness of your results.

Towards the end of the Simulation section, We added “ The surface climate responses are evaluated based on the 20-year average over the period of 2050-2069. With three ensemble members, the 20-year average of each evaluated climate variable is calculated based on 60 annual-mean values. Taking into account temporal autocorrelation, the effective sample size is still comparable to the suggested number of independent data points (20-40) in Pausata et al. (2015). This effective sample size is also comparable to the suggested sample size (7-40) in another relevant study focusing on discerning NH polar vortex change from internal variability (Bittner et al., 2016). As this study focuses on the long-term impacts of continuous injection, rather than impacts of a pulse volcanic eruption in the single year following the eruption (as in, e.g., Pausata et al., 2015; Bittner et al., 2016), data from three ensemble members are likely sufficient to distinguish a signal over a 20-year period from internal variability. ”

2. In the abstract, you could consider removing or shortening the following passage: “therefore, understanding the range of possible climate outcomes is crucial to making informed future decisions on SAI, along with the consideration of other factors. Yet to date, there has been no systematic exploration of a broad range of SAI strategies. This limits the ability to determine which effects are robust across different strategies and which depend on specific injection choices, or to determine if there are underlying trade-offs between different climate goals.” This reads more like a passage from the introduction, justifying the novelty of the article, than an abstract. I however accept that this point is to some extent subjective.

We have shortened this passage by removing the following part, “, or to determine if there are underlying trade-offs between different climate goals”.

3. As Wilks (2016) points out, multiple testing issues can lead to spurious spatial patterns of significance, which are impossible to evaluate a priori and which can in turn lead to incorrect interpretation of results. Since you cite Wilks and are clearly aware of the issue, simply mentioning it in a sentence but not correcting your figures is hard to justify scientifically. This is particularly relevant for the global geographical plots. I would warmly encourage you to update your testing for these. There are ready-to-use packages performing multiple testing correction in several programming languages, which cause only a modest computational overhead.

We thank the editor for their comment. We have performed multiple testing on regional changes in temperature, precipitation, and P-E to account for spatial correlation, and updated the t-test results in Figures 8-10 in the manuscript and Figures S2-S6 in the Supplement. In Line 168 in the manuscript, we added “We also perform multiple testing to account for spatial correlation using the false discovery rate (FDR) method, where we choose α_{FDR}=0.1 for achieving a global significance level of 0.05 based on the conclusion in Wilks (2016).”