Summary:

This paper assesses the potential for stratospheric aerosol injection (SAI) to reduce permafrost thaw using Earth-system model simulations run under the G6sulfur and G6solar protocol, which are a part of the broader Geoengineering Model Intercomparison Project (GeoMIP). The output from the G6sulfur and G6solar simulations are compared to ssp585, which serves as the background climate change scenario, as well as ssp245, which serves as the target of SAI deployment in each of these SAI deployment scenarios. SAI deployment in both the G6sulfur and G6solar simulations is found to reduce the rate of decline of permafrost area and thaw depth expected under ssp585 to a rate that is more comparable to that expected under ssp245. However, the rate of decline in permafrost area thaw depth in G6sulfur and G6solar is found to be larger than that in ssp245. Variance partitioning and correlation analyses are used to identify physical mechanisms that may be driving the identified discrepancies. The discrepancies between the G6sulfur and ssp245 simulations are connected to changes in atmospheric circulation in both models, suggesting that some of the discrepancies between these two scenarios may be due to a more positive NAO signal during the winter season in G6sulfur.

While the focus of this paper, to increase understanding of the potential impact of SAI on permafrost thaw, is an interesting research question, some aspects of the background and methods are unclear and key arguments that are made are not fully convincing. For one, there is not enough discussion and description about permafrost and related processes (e.g., permafrost stability, explanation of permafrost as a tipping element, permafrost as a carbon store) to allow the manuscript to be understood by a general audience. Second, the authors should consider reframing the motivation behind studying the potential risks and benefits of SAI. The key methodological choices that need to be addressed and/or clarified are related to the statistical significance methods considered, the handling of different numbers of ensemble members for each simulation of considered models, and an inadequate description of the variance partitioning analysis. The argument that permafrost area and permafrost thaw depth in G6sulfur and G6solar are significantly different from ssp245 is not totally convincing given that there is not a full description of the details of the significance testing including what are considered to be independent samples and what the degrees of freedom are, insufficient display of uncertainty, and errors in how the multimodel means are computed. Strengthening this argument is key for the remainder of the manuscript's analysis, which investigates why the differences between the SAI deployment simulations and ssp245 exist. More detailed major and minor comments are outlined below:

Major Comments:

- A major point of the results is that while the deployment of SAI in G6sulfur and G6solar are able to reduce the rate of decline of permafrost area and thaw depth compared to ssp585, the SAI deployment simulations rates of decline are still larger compared to ssp245, which is the considered to be the target of deployment (Figures 1 and 2). The authors test for significance using the two-sample Welch's t-test and state that the

difference between the G6sulfur and G6solar simulations and ssp245 for the 2080-2099 period are statistically significantly different from ssp245 for permafrost area and thaw depth. The time series line (Figure 1a) and the spatial plots shown in Figure 2 in combination with the statistical significance testing is not a fully convincing argument that the results from the G6sulfur and G6solar simulations are significantly different from that in ssp245. To attempt to strengthen this argument, the authors should first provide more information about the details of the significance testing in the methods section (i.e., what is considered to be an independent sample (each model, each member?) and what the number of degrees of freedom are). Additionally, the authors should include the range of values predicted by individual ensemble members as uncertainty bars in plots such as Figure 1a (as is shown in Figure 4). The authors should also consider providing some measure of uncertainty related to the permafrost thaw depth difference plots shown in Figure 2a.

- The authors should consider reframing the introduction and include more details on the concepts surrounding permafrost including why permafrost acts as a carbon store and the transition of permafrost from a carbon store to a carbon source. This discussion should also explain why permafrost is a tipping element, and what implications permafrost collapse would have. Additionally, the authors should rewrite the introduction of the motivation to research the potential risks and benefits of methods of climate intervention, including SAI. While the critical role of Arctic permafrost and the risks posed by its thaw do provide motivation for research into the potential impacts of SAI, it is not the only motivator. The authors might point out that global mean temperature rise is projected to exceed 1.5^{o}C in the next decade (IPCC, 2022), and then include a bit of discussion of some of the expected implications of exceeding this level of global temperature rise. The authors should consider removing discussion of small-scale field experiments that were cancelled, as well as private companies attempting to deploy SAI, as it is not relevant for discussion in this manuscript.
- There are a wide range of ensemble members used from each respective model (e.g., the number of members for CESM2-WACCM under ssp245 is 4-5, and the number of members for MPI-ESM1-2-LR is 50). Thus, just taking the mean across all members and models gives more weight to models that have more members. Further, for each simulation, the number of ensemble members available for each model varies. The authors should weight ensemble means so that each model (e.g., CESM2-WACCM, CNRM-ESM2-1, etc.) has an equal contribution to the ensemble mean so that each model contributes equally to the multimodel mean. Additionally, Table 1 shows that for some simulations, the number of ensemble members for some models varies (e.g., for CESM2-WACCM ssp245 simulations, 4-5 ensemble members are listed). I assume that this means that not all members of each simulation output each variable that is considered. Somewhere in the manuscript (e.g., Table 2, or in the supplement) what number of ensembles is available for each simulation in each model for each variable

should be listed explicitly. The weighting by number of ensemble members for each simulation for each model should then be adjusted accordingly.

- What is actually done in order to partition the variance contribution from each of the considered potential predictors is not explained with adequate detail. First, why are models constructed using only three predictors? What is the sensitivity of the results shown to the number of chosen predictors? LMG needs to be written out explicitly. What LMG is/does also needs to be defined and discussed. What does it mean to construct "all possible models comprising three predictors"? Why does the order of predictors in each model matter? How sensitive is this variance partitioning analysis to potential relationships (correlation) between predictors? In addition to adequate discussion of the details of the variance partitioning analysis, the interpretation of the results of this analysis should also be clarified. Some specific sections that need clarification are listed below:
    - Lines 345-351: Why might downwelling longwave radiation, near-surface air temperature, and northward wind at 20 hPa explain the largest portion of the variance of the difference between the G6sulfur and ssp245 simulations?
    - Lines 352-357: Why are the variables which describe the largest portion of the variance entirely different between the winter and summer, when both are to predict SPTD?
    - Lines 386-393: This discussion hinges on the results shown in Figure 4. While there are subtle differences between the ensemble means of G6sulfur and ssp245, the range of values are overlapping in all cases, suggesting that none of the differences discussed in this section are significant. The authors might consider re-doing Figure 4, showing the G6sulfur-ssp245 for each individual model as its own line. Then, deviations from zero can be discussed.
    - Figures 3c, 3d, 5c, 5d, and related subplots in Supplementary Figures 9 and 10 contain way too much information, most of which is unused. In section 3.3 specifically, the only correlations that are discussed are those between thaw depth (td) and the 18 predictors. The authors should revise these figures to only include the relevant correlations (the first column). There is no need to include the correlations of each predictor with the other predictors when those correlations are not included in any interpretation of the results.
- The discussion connecting the NAO to differences in permafrost area and thaw depth between G6sulfur and ssp245 is not convincing as written.

Minor Comments:

- The Abstract should be edited to include mention of the G6solar simulations, which are not currently mentioned.
- Please make sure that each citation has a publication date associated with it, there are currently a few that include no date.

- In plots such as Figure 6a-6c, the authors should consider statistical significance testing by grid point, adjusted for the False Discovery Rate (Wilks, 2016), to show regions where the G6sulfur simulations differ significantly from ssp245, rather than field significance tests.
- Lines 86-88: Add citation (e.g., Crutzen, 2006)
- Lines 84-85: Rephrase "strategies to manipulate the climate to reduce warming" to something like "strategies to prevent global temperatures from surpassing dangerous thresholds"
- Lines 169-170: While Kravitz et al. (2015) states that the G6sulfur scenario was designed based off of radiative forcing reduction goal, Jones et al. (2022) states that this was eventually updated to a global mean near-surface air temperature goal. The manuscript should be updated to reflect this change.
- Lines 172-174: cite O'Neill et al. (2017)
- Lines 176-182: This paragraph should be updated. The amount of aerosol or aerosol precursors injected in each model in G6sulfur is adjusted such that global mean near-surface air temperature is within 0.2ºC of the global mean near-surface air temperature in ssp2-4.5, rather than a reduction in radiative forcing. It should also be noted that some of the models could not directly simulate the injection of aerosols into the stratosphere for the G6sulfur experiment, and so instead prescribe AOD distributions provided by GeoMIP (CNRM-ESM2-1) or by Neimeier and Schmidt (2017) and Niemeier et al. (2020) (MPI-ESM1-2-LR and MPI-ESM1-2-HR) (see Jones et al. (2022)).
    - Lines 591-594 should also be updated accordingly.
- Lines 213-314: What is the complementary error function? Is it used to diagnose uncertainty? This explanation should be expanded.
- Lines 216-218: Why are only grid cells with a land fraction greater than or equal to 35% and grounded ice sheet fraction less than or equal to 65% included? Has this been done before in similar studies? Include justification and citations here.
- Lines 222-223: Which models might have better (or worse) resolved soil temperature profiles?
- Line 305: Why is thaw depth a critical element of permafrost stability?
- Line 463: What is the NAO index?
- Lines 594-596: The sentence should be rephrased to say something like, "Although rates of decline of permafrost area and thaw depth are slowed in G6sulfur and G6solar relative to ssp585, permafrost area and thaw depth declines more rapidly under SAI than in ssp245." This sentence should only remain in the manuscript if this result is valid after addressing the aforementioned methodological concerns.

Citations

Crutzen, P. J. (2006). Albedo Enhancement by Stratospheric Sulfur Injections: A Contribution to Resolve a Policy Dilemma? *Climatic Change*, *77*(3), 211–220. https://doi.org/10.1007/s10584-006-9101-y

IPCC. (2022). *Global Warming of 1.5°C: IPCC Special Report on Impacts of Global Warming of 1.5°C above Pre-industrial Levels in Context of Strengthening Response to Climate Change, Sustainable Development, and Efforts to Eradicate Poverty* (1st ed.). Cambridge University Press. https://doi.org/10.1017/9781009157940

Jones, A., Haywood, J. M., Scaife, A. A., Boucher, O., Henry, M., Kravitz, B., et al. (2022). The impact of stratospheric aerosol intervention on the North Atlantic and Quasi-Biennial Oscillations in the Geoengineering Model Intercomparison Project (GeoMIP) G6sulfur experiment. *Atmospheric Chemistry and Physics*, *22*(5), 2999–3016. https://doi.org/10.5194/acp-22-2999-2022

Kravitz, B., Robock, A., Tilmes, S., Boucher, O., English, J. M., Irvine, P. J., et al. (2015). The Geoengineering Model Intercomparison Project Phase 6 (GeoMIP6): simulation design and preliminary results. *Geoscientific Model Development*, *8*(10), 3379–3392. https://doi.org/10.5194/gmd-8-3379-2015

O'Neill, B. C., Kriegler, E., Ebi, K. L., Kemp-Benedict, E., Riahi, K., Rothman, D. S., et al. (2017). The roads ahead: Narratives for shared socioeconomic pathways describing world futures in the 21st century. *Global Environmental Change*, *42*, 169–180. https://doi.org/10.1016/j.gloenvcha.2015.01.004

Wilks, D. S. (2016). "The Stippling Shows Statistically Significant Grid Points": How Research Results are Routinely Overstated and Overinterpreted, and What to Do about It. *Bulletin of the American Meteorological Society*, *97*(12), 2263–2273. https://doi.org/10.1175/BAMS-D-15-00267.1