Reviewers' comments are in bold. Authors' responses are in blue.

General comment

This study uses a multi-model ensemble of global aerosol simulations performed within ISA-MIP HErSEA to assess the effect on volcanic stratospheric aerosol of uncertainties related to the SO_2 injection (height and amount) by the 1991 Pinatubo eruption. As a main result, the study identifies large inter-model differences as well as common limitations, particularly related to a too strong simulated meridional transport of aerosol in the northern hemisphere, that results in a faster simulated decay of the post-eruption enhancement of the stratospheric aerosol layer compared to observations. The study also highlights how different SO_2 injections are required for different models to "best match" observations (and how these vary for the chosen observed parameter as well).

I have only minor comments on the study, which I found overall well-conceived and well conducted. My evaluation of the study considers it as a "MIP" study, so based on results from a predefined protocol-driven set of experiments. I recognize that some aspects of the study remain open to discussion and thus require further investigation (the role of the Cerro Hudson and the role of ash emission as far as comparison with observations is concerned, but also the causes of the found inter-model differences). This calls for a retrospective on the HErSEA protocol (was it effective or has any weakness emerged?) and for a discussion about the implications of the findings for the original purpose of the experiment and for the purpose of ISA-MIP in general (this is mentioned for instance in lines 61-62 of the manuscript). As another general comment on the study, I encourage a more explicit discussion (if not presentation) of within-model uncertainties, intended as differences between realizations of an experiment with the same model. These might be negligible in most cases, but this is not stated and, instead, there are occasions where illustration of results from individual realizations reveals distinct behaviors (for instance in Figure 3). I have some more specific comments on this below.

I have also just a few minor editorial comments, as in my opinion the manuscript is overall well-structured and well written. As a general comment, I felt there is a difference in style between sections 3.1 and 3.2 (just focused on presentation of results) and section 3.2 (which mixes introduction, results and discussion, especially from the paragraph starting on line 374 onward). Maybe the authors could consider some homogenization, for instance by moving some of the more discussive parts of section 3.2 in section 4.

We thank the reviewer for his suggestion. We moved most of the discussion of section 3.2 in the discussion section, changing much of its structure.

Then, the manuscript could serve as a reference for future analyses based on the HErSEA experiments, especially as far as final choices in the experiment setup differ from the original protocol. In this sense, it may be worth to provide any guideline provided for the generation of the ensemble, and how this was actually done for each model. I see that for most models this is not reported, while in the other cases it is

not clear if the parameter perturbation was maintained for the whole simulation or just for some initial steps (ECHAM6-SALSA).

We specified in the experimental protocol section that “The generation of the ensemble for each model is explained in the respective sections describing the model.” and we did as mentioned. In particular for ECHAM6-SALSA we specified that “Ensemble members were produced by using insignificantly different values for one of the tuning parameters (the rate of snow formation by aggregation) for January 1991 of each ensemble member.”

Specific comments

Line 44-46: maybe it is worth mentioning here that a possible cause of the inter-model discrepancies in radiative fluxes are minor differences in forcing implementation.

We have revised the paragraph to make it more clear. As we focus in our study on the comparison of global interactive aerosol models we will refer now only to VolMIP wrt to the Tambora study as a VolMIP pre-experiment and do not discuss VolMIP results in general.

Line 58: proposed cooling is unclear, maybe “a certain cooling target”?

We specified the proposed cooling target in order to be clear, as follows:

“The Geoengineering Model Intercomparison Project Phase 6 (GeoMIP6, Kravitz et al., 2015) also includes experiments with injection of stratospheric sulfate aerosols precursors in an amount to reduce the net radiative forcing from the SSP5-8.5 scenario to the SSP2-4.5 scenario ”

Line 61: to me initial conditions refer to the initial state of the system as a whole, so more than the “initial conditions of SO_2 injection” that is implicated here. I recommend the authors to always explicit this to avoid confusion. Also, other “initial conditions” such as the phase and amplitude of the QBO may be relevant here and deserve some explicit consideration in the presentation and discussion of results (see also comments below).

We specified that initial conditions refer to the different SO2 injection settings and defined in section 2.1.1 (Experimental Protocol) the implementation of QBO, which is discussed in the results section.

Line 161: by climatological do you mean “observed” values during the simulated period?

We changed “climatological” in “observed”.

Line 267: is this related to the QBO phase? There seem to be little information regarding this aspect in the presentation of results and discussion. If the model spontaneously produces a QBO, it would be instructive to know how QBO phase and amplitude compare with observations. In this regard, one of the realizations of ECHAM6-SALSA is clearly different from the other two, especially in terms of rms (see Figure 3): what is the reason behind this difference? I wonder if the ensemble

mean is truly representative for this model at least. This might motivate some focus on individual realizations as well (or on sub-ensembles).

We have now discussed in the paper the details of the experimental protocol, which prescribed a QBO consistent with observations, also for models with interactive QBO (that needed to control for consistency in their QBO state). Therefore, the observed intra-ensemble differences can't be due to different QBO states. We have added the following phrase:

"The evolution of the quasi-biennal oscillation (QBO) must be consistent through the post-eruption period, as it affects the dispersion of the volcanic plume to mid-latitudes (Trepte and Hitchman, 1992; Baldwin et al.; Punge et al., 2009), and consequently the size distribution and lifetime of stratospheric aerosols (Hommel et al., 2015; Pitari et al., 2016; Visioni et al., 2017). Accordingly, models with internally generated QBO re-initialized it in order to be consistent with the actual meteorological conditions, or used specified dynamics approaches (e.g. Telford et al., 2008)."

Line 354: why not testing the differences? Even if the sample size is low, a Mann-Whitney U test, for instance, could provide you a basis for a stronger statement here.

We prefer to show that the differences between the ensemble members of the same scenarios in ECHAM6-SALSA can be larger than the the differences between the ensemble mean of different scenarios as in Figure 1 (S1 in the supplementary material): the thick line represent the ensemble mean of each scenario and the shaded area the region between the minimum and maximum value between the ensemble members.

We added this figure in the supplementary material and referred to it in that paragraph.

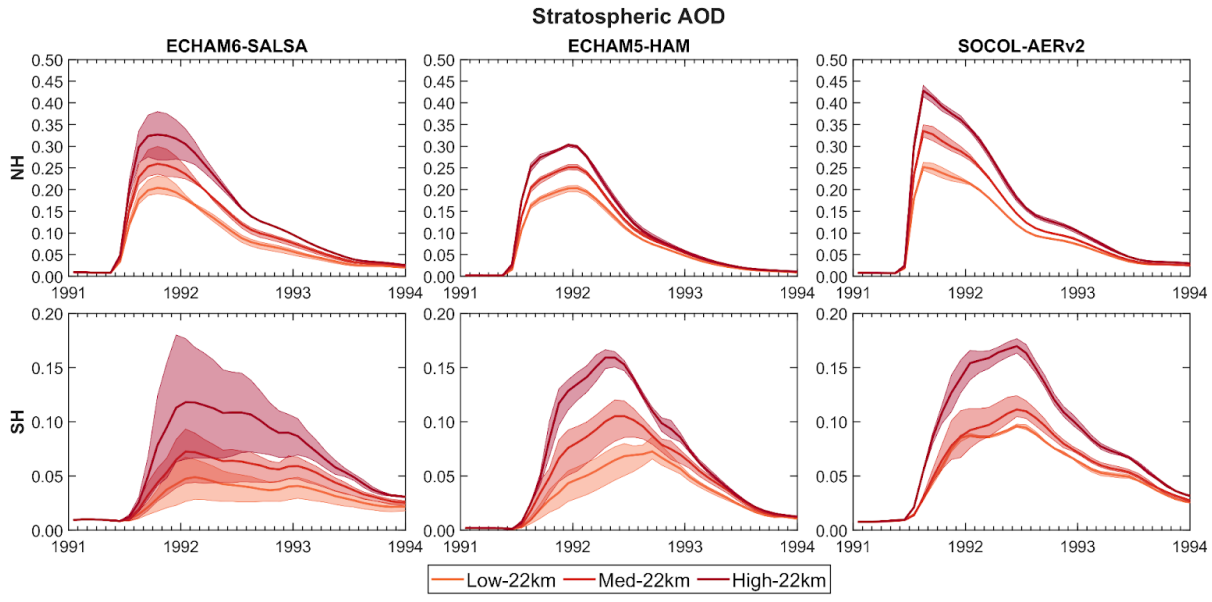

Figure 1. Time evolution of the stratospheric AOD in the northern (NH) and southern hemisphere (SH) simulated by ECHAM6-SALSA, ECHAM5-HAM and SOCOL-AERv2 for the experiments with different masses of SO2 injected at about 22 km altitude. The thick

line represents the ensemble mean, the shaded area the region between the minimum and maximum value between the ensemble members (thin lines).

Figure 8: especially for the Laramie comparison, given the punctual location of the datum, would it make sense to consider more explicitly the individual realizations instead of just the ensemble mean in order to include uncertainties linked to the "internal component" of atmospheric circulation? I understand that also due to the vertical averaging this might still lead to small differences across realizations, but it would be important to have some estimate of the uncertainty anyway (for instance an error bar at the peak value of the profile). Also, the error bar for the OPC data is not defined.

At the beginning of section 3.3 we refer to Appendix A2 for all calculations related to the effective radius and error bar. We find that adding the shaded areas (that represent the area between the minimum and maximum values of the three ensemble members) makes the figure too messy for ECHAM6-SALSA and doesn't add any further information that has not already been discussed.

Technical corrections

Line 332: typo (produces) Corrected.

Line 391: twice especially, maybe the second can be skipped Corrected.

Line 425: at analysing Corrected.

Line 574: typo Higher. Corrected.

Figure 3: I had some difficulties tracking the colors. I suggest using a more varied color palette for the different experiments. We changed the colors using a diverging scheme ("RdYlBu") for which we made sure that was colorblind safe. The same palette is used for the comparison of experiments, with the exception of the figures where experiments of all models are compared at the same time (Figures 7, S1, S2, S7). In that case, we left the different linestyle for the experiment, as specified in the caption ("Experiments are identified here with different line styles, the different colors refer to the models.")

References

Baldwin, M. P., Gray, L. J., Dunkerton, T. J., Hamilton, K., Haynes, P. H., Randel, W. J., Holton, J. R., Alexander, M. J., Hirota, I., Horinouchi, T., Jones, D. B. A., Kinnersley, J. S., Marquardt, C., Sato, K., and Takahashi, M.: The quasi-biennial oscillation, Reviews of Geophysics, 39, 179–229, https://doi.org/https://doi.org/10.1029/1999RG000073, 2001.

Hommel, R., Timmreck, C., Giorgetta, M. A., and Graf, H. F.: Quasi-biennial oscillation of the tropical stratospheric aerosol layer, Atmospheric Chemistry and Physics, 15, 5557–5584, https://doi.org/10.5194/acp-15-5557-2015, 2015.

Kravitz, B., Robock, A., Tilmes, S., Boucher, O., English, J. M., Irvine, P. J., Jones, A., Lawrence, M. G., MacCracken, M., Muri, H., Moore, J. C., Niemeier, U., Phipps, S. J., Sillmann, J., Storelvmo, T., Wang, H., and Watanabe, S.: The Geoengineering Model Intercomparison Project Phase 6 (GeoMIP6): simulation design and preliminary results, Geoscientific Model Development, 8, 3379–3392, https://doi.org/10.5194/gmd-8-3379-2015, 2015.

Pitari, G., Visioni, D., Mancini, E., Cionni, I., Di Genova, G., and Gandolfi, I.: Sulfate Aerosols from Non-Explosive Volcanoes: Chemical-Radiative Effects in the Troposphere and Lower Stratosphere, Atmosphere, 7, 85, https://doi.org/10.3390/atmos7070085, 2016.

Punge, H. J., Konopka, P., Giorgetta, M. A., and Müller, R.: Effects of the quasi-biennial oscillation on low-latitude transport in the stratosphere derived from trajectory calculations, Journal of Geophysical Research: Atmospheres, 114, https://doi.org/https://doi.org/10.1029/2008JD010518, 2009.

Telford, P. J., Braesicke, P., Morgenstern, O., and Pyle, J. A.: Technical Note: Description and assessment of a nudged version of the new dynamics Unified Model, Atmospheric Chemistry and Physics, 8, 1701–1712, https://doi.org/10.5194/acp-8-1701-2008, 2008.

Trepte, C. R. and Hitchman, M. H.: Tropical stratospheric circulation deduced from satellite aerosol data, Nature, 355, 626–628, https://doi.org/10.1038/355626a0, 1992.

Visioni, D., Pitari, G., Aquila, V., Tilmes, S., Cionni, I., Di Genova, G., and Mancini, E.: Sulfate geoengineering impact on methane transport and lifetime: results from the Geoengineering Model Intercomparison Project (GeoMIP), Atmospheric Chemistry and Physics, 17, 11 209–11 226, https://doi.org/10.5194/acp-17-11209-2017, 2017.