By Larry Thomason

I marked this as requiring major revision but it is really somewhere between minor and major.

In general, this is a detailed paper that is well written and very helpful understanding the CMIP stratospheric aerosol forcing paper. At times, it is so detailed that I found it kind of difficult to slog my way through it but that is OK. However, at times I find perhaps even the authors found it hard to go into the required detail and there are too many blanket statements that lack required detail (or a reference) to support the assertions made. In addition, I am concerned by the number of comparisons, and the conclusions drawn there from, with hemispheric and/or multi-decadal averaging sometimes compared to point measurements. To me, it presents the possibility that interesting components of the comparisons could be lost or erroneous conclusions could be made. I find that particularly worrisome. The paper could use a score card with all the eruptions mentioned in the paper with a reference.

Line 33, What is an intermediate complexity aerosol model? Also see line 104 for 'reduced complexity model'.

Line 34-35. It would be nice if there was a generally accepted definition of what constitutes a small, moderate, or large SO2 event. This isn't the place to solve and at least the authors define some categorization. However, even here, does 0.3 to 3.0 Tg comprise small and moderate eruptions (so what are they below 0.3 Tg?) or do they really mean moderate? Certainly people throw around qualitative words like these without defining them and it is meaningless.

Line 192. What does this statement add?

Lines 204-213. I wonder, and perhaps the references cover this question, but how certain can the record with time be if say, observers only report the SAOD (or stellar magnitude) when there's something interesting to report as opposed to whenever they make a measurement. If there's some sort of optimal measurement selection, it may be possible to bias average values toward large values if only large values are reported. I don't know the answer here.

Line 215-237. It might be worthwhile to mention airborne lidar data from NASA missions in the 1980s following El Chichón and the 1990s following Pinatubo here. Thomason et al (2018) archive these data at the NASA Atmospheric Sciences Data Center.

Figure 2. For all the text about lunar occultation data, it really looks awful.

Line 288-289. With the range of values possible I am not sure how informative global average differences are.

Section 3.1.2. Given the large range of optical depths, I wonder if large events would dominate the observed overall structure shown in this figure/text.

Line 335-339. These are relatively short periods of time in terms of episodic events like volcanoes (and the satellite data in GloSSAC and others include smoke events at a mostly low level). I suspect that if you divided the satellite period in a favorable manner you could produce differences as well. Sort of like the 'solar cycle' signal in SAOD depth between El Chichón and Pinatubo, some things are artifacts of timing. So, I find saying 'likely' is a bit strong, and 'possibly' would be better.

Line 345-348. I am not real enthusiastic about any of these 'background periods. I am pretty sure the stratosphere, at least at some altitudes/latitudes is still relaxing back toward a 'background' until at least 1999 if not 2000. As the authors point out 2002 had at least 1 significant event (Ruang) and 2005 had an eruption by Manam in January. As I recall, GloSSAC at least version 1. Had a background constructed from 2000, 2001, 2003 and 2004 but even there, the was a small but clear trend from January to December particularly in the tropics. (Confession I did this work) I may, but it is getting fuzzy, have done something to mitigate the trend but I do not recall now. So, it would be nice to see some comparisons of these 'backgrounds' to see whether there is a residual trend still in these values. I assume under optimal circumstances the authors would like there to be no trend.

Line 348-352. So CMIP 7 is higher than 6 only due to the change in years used in computing a background (and particularly 1998 which is clearly not post Pinatubo)? If so, where does the trend in the pre-satellite data come from? There is no obvious discussion as to how this could occur. Linear trend vs not linear? Is there a reason to prefer kind of trend over the other. This is under-discussed.

Figure 5. Wow, the differences between events can be huge. Many of the comparisons don't even look like the same events and it is worth saying so right up front in this discussion. Some explain of what causes the huge differences would be helpful. I would also like to see relative differences in addition to absolute differences. I suspect that the relative differences far exceed a factor of 10 at times between CMIP6 and 7 at important levels of extinction coefficient. Could the color bars be made wider? It's hard to use it to infer values in the plots.

Figure 6. If you are going to compare CMIP& data to a single pyrheliometer then the data from the models should match the latitude and not use a global mean. As it is, the comparison is meaningless.

Figures 7-10. There is some information missing from the discussion and the captions for these figures. What are there two line for CMIP6 and CMIP7? It's rather confusing.

Line 408 and following. While I know that GloSSAC needs some updating particularly in the northern hemisphere to account for the lack of data (which is underway), I worry that assuming a point measurement is representative even in it's own latitude region a short time after an eruption. Things can remain quite inhomogeneous for some time. See Winker's 1991 paper on airborne lidar measurements of the Pinatubo cloud for an example. Applying that point measurement to a huge latitude range is even more questionable. At a minimum, for point measurements like Mauna Loa, comparisons should be made with appropriate latitude zonal mean data like GloSSAC.

Line 431, Is 25% 'slightly' lower?

Line 490. What are these volcanically quiescent periods? They are pretty hard to come by.

Line 480-481 and following. I know that CMIP is more or less forced to assume things about aerosol size distributions but I am worried that information for the occultation period suggests some particularly not-log normal or at least not typical size distributions may occur following even small events. (sorry for self-referencing) See for instance, Thomason et al 2021 (https://doi.org/10.5194/acp-21-1143-2021). How big a deal is the assumed size distribution to the end products that CMIP7 requires?

Figure 14. There's some really big averaging across huge changes in the stratospheric aerosol loading in figures d-g. I am not sure how to interpret this figures with things change so much in time.

Figure 17. The number densities are labeled as 10^9 molecules per cm^3 . I don't think this can be right. Aerosol number densities are usually something like 5 or 10 cm^{-3}. Also, as is well known, lower moments of inferred aerosol size distributions are poorly contained in VIS-NIR aerosol extinction coefficient measurements and the results of this comparison are going to be totally model dependent rather than an actual inference of a meaningful number density. I hope the authors are aware of this and make it clearer to the reader.

I really liked the discussion in section 4.1. Too many people assume that VEI tells the whole story. Overall, Section 4 is very illuminating.

Lines 636-689. It would be interesting (somewhere) to see the difference in the aerosol wavelength dependence between SAGE II and Sciamachy. That dependence is what is likely contributing to the effective radius differences (and the model).