We thank the reviewers again for making very useful suggestions to further improve the paper. Our point-by-point responses to the reviewers' comments and corresponding changes are detailed below in blue text, and the changes are shown in the version of the manuscript with track changes.

Report #1
Anonymous Referee #1

I appreciate the work the authors have put into improving the manuscript. The method of obtaining SO2 values from extinction observations is now less opaque than it originally was, and many smaller details have been improved. I do however still have some concerns with the manuscript. I will here revisit the major comments from my first review, then move into specific comments on the revised manuscript.

General comments:
1. On the method of estimating SO2 injection from satellite aerosol extinction observations, I do appreciate the new material added, which help to understand the method. But still many important details are missing. I point out that according to the ACP review criteria, my job is to ask "Is the description of experiments and calculations sufficiently complete and precise to allow their reproduction by fellow scientists (traceability of results)?" Presently, I do not think this is the case. Here are some unresolved questions I have about the method:

• The method uses data typically from "a ten-day period beginning about a week after the eruption". However, the given e-folding lifetime of SO2 is between 13 and 32 days according to Hoepfner et al. This suggests that at the period from which the extinction measurements are used, not all SO2 has been converted to sulfate aerosol, by an amount that depends on altitude. Is it assumed that all SO2 has been converted, and if so, can you quantify the impact this assumption would have on your results?
-> Response: In section 4 is added "... in the tropics. For higher latitudes the selected period is later and longer, taking into account the longer conversion time (due to less OH)." More on that in Appendix B and the electronic supplement.
• The effect described above could be exacerbated by the fact that satellite instruments may be insensitive to aerosol particles of very small size. Some related discussion would be useful.
-> Response: We always check the total amount of injected SO2 derived by our method against estimates from nadir instruments for consistency as stated in the text. Volcanic aerosol and its extinction is always dominated by the accumulation mode with a small fraction from the Aitken mode (and bigger particles for Pinatubo).
Satellite measurements are not insensitive to aerosol extinction by very small particles, but it is true that the retrieval of aerosol extinction by very small particles is challenging because such particles result in Rayleigh scattering that has to be distinguished from Rayleigh scattering by neutral air density. For this reason, the estimation of this contribution from neutral air has to be as accurate as possible. We added as short discussion on this issue, and more detail about the GOMOS retrieval in section 2.2.
• A key part of the method is the construction of a scaling factor from model results between extinction and sulfate mixing ratio. It is not obvious that a single number will be universally applicable to this purpose, rather one would expect the scaling factor would depend (perhaps strongly) on the aerosol size distribution. I think a detailed illustration of the extraction of this scaling factor from model results and a discussion of the uncertainty in this scaling factor is essential to the description of this novel method.
-> Response: This is explained in more detail now in the main text and Appendix B. The uncertainty depends on the individual eruptions and the data coverage but it should be less than the one from the idealized assumptions on the vertical distributions in the point source methods.
• At line 313 of the tracked changes document, can you explain why one needs to divide by air density? If the scaling factor allows one to covert from extinction to sulfate mixing ratio, its not clear to me why air density is needed to get to SO2 mixing ratio.
-> Response: We have modified the text concerning that to be more precise. The extinction is proportional to the concentration so to get a mixing ratio you have to divide by density as also shown in the formula in Appendix B.
• More detail on how the spatial structure of the extinction plume is quantified is needed. There are many references to the 3D structure of the plume, however, references to the zonal mean are also present (e.g., line 317), and so it's not quite clear if the quantities are based on full zonal averages or an integration over a limited spatial range. And if the latter, how that range is defined. It's also not clear to me if the method assumes that the aerosol plume is stationary in space. If the plume is moving, then over the 5 day period of measurements, similar parts of the plume may be measured in different locations. This would seem to be a significant source of uncertainty in the method.

-> Response: It is not our objective to model the details of the plume or local effects but the impact of the volcanoes on climate forcing. There is some smearing out of details due to the use of data averaged over a period of several days. The examples with point sources in Appendix C show that the forcing can be similar but that the uncertainties from the assumptions on the vertical distribution are larger. The text on spatial integration at the beginning of section 5 has been corrected (see also details).

• There is a final correction factor which takes values between 0.5 and 3 (so, close to an order of magnitude of impact on the final result) to correct for time gaps between SO2 injection and extinction measurement, or the influence of prior eruptions. The description of how this correction factor is derived and applied is inadequate. What data sets are used to construct this correction factor, and how exactly does it depend on time since eruption and any other factors (latitude, injection height, etc). I assume that the correction factor is constructed from cases where both extinction and MIPAS SO2 measurements are available, and then it is applied in the cases when MIPAS measurements are not available, but this is not explicitly stated anywhere. Please elaborate.

-> Response: We hope we have clarified the concept of the correction factor now in the text. As described in more detail in Appendix B the correction factor is always case dependent. If the time lag of data is several weeks a correction factor >1 has to be applied to account for removal processes, if another event is relatively close in time, the factor has to be <1 to remove the influence of the previous event. Factor 1 is default. For convenience, we supply now a table with the correction factors and the selected times for OSIRIS in the electronic supplement (referred to in section 4).

• I do not understand the advantage of using SAGE II SAD values rather than extinction observations for the analysis. SAD is a derived product, which comes from the extinction but incorporates a number of assumptions, and is therefore more uncertain than the extinction measurements. Plus, since the study uses extinction measurements from OSIRIS and GOMOS, it would be more consistent to use the extinction measurements from SAGE II. The justification given for using SAGE II SAD is that it "it is easier to correct for cloud contamination", but why exactly is this?

-> Response: We used both approaches as shown in Appendix B.

In the lower part SAD has a slightly better data coverage than the extinction at 525 nm. Another reason for using SAD here is for comparison of results with our earlier studies.

2. Concerning the table of results, the formatting is improved compared to the first version. My opinion is that a format with one event per line would be much easier for users to use (it could be read into a spreadsheet program for instance), the FORTRAN formatted table would be difficult for some users to parse. But I leave this to the discretion of the authors and editor.

-> Response: Comprehensive models normally use netcdf or ascii data as input as provided in the link. Preferably the 3D-data should be used, the table is more for comparison purposes as indicated at the beginning of section 5. A table with about 550 lines would be unacceptable for the readership. The caption of Table 2 is slightly modified to be more clear what is available in the link.

3. Finally, concerning the agreement between the model output and observations, I appreciate the work the authors have done to adjust the colorbars of the various figure to make them consistent. This makes quantitative comparison possible for the reader, if not especially easy. Comparing the tropical extinction profile timeseries between the model (Fig 7, 8) with the SAGE II (Fig 5), OSIRIS (Fig 4) and GOMOS (Fig 3) extinctions, there seems to be a noticeable difference in the vertical profile, with the model results showing larger extinctions generally between 18-20 km and lower extinction below 16 km. This difference seems to cancel out in the SAOD, but should not be ignored.

-> Response: When comparing the vertical distribution of aerosol extinction at 750 nm between the satellite observations below 16 km altitude of GOMOS, OSIRIS and SAGEII one recognizes the different sensitivities of the instruments on cloud perturbations and measurement uncertainties. As written in chapter 2:

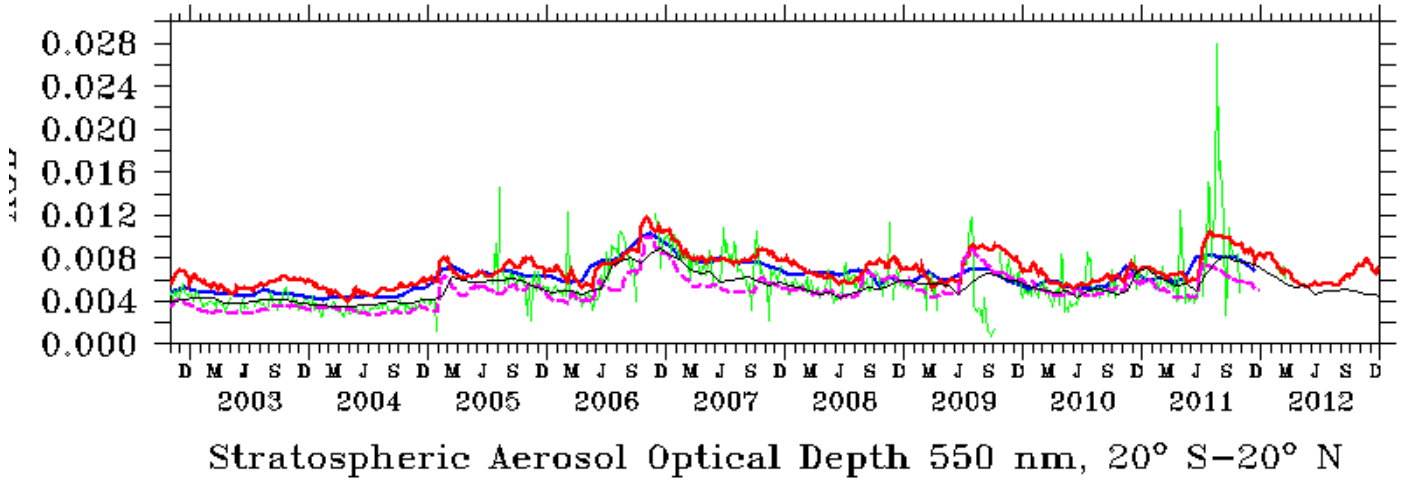
OSIRIS: "At altitudes near and below the tropopause, the OSIRIS measurements are sensitive to clouds that may be interpreted as elevated aerosols. This is likely contributing to larger background extinction values measured below approximately 17 km in the tropics, as can be seen in Figure 4 (bottom), and the uncertainty is higher."

SAGE: "Red pixels around 14–16 km correspond to measurements contaminated by clouds, increasing the optical depth in the upper troposphere/lower stratosphere (UTLS) region on the lower panel of Figure 5. The perturbations by convective clouds occur mostly over the West Pacific and were excluded in the procedure for estimating the SO2 injections. "

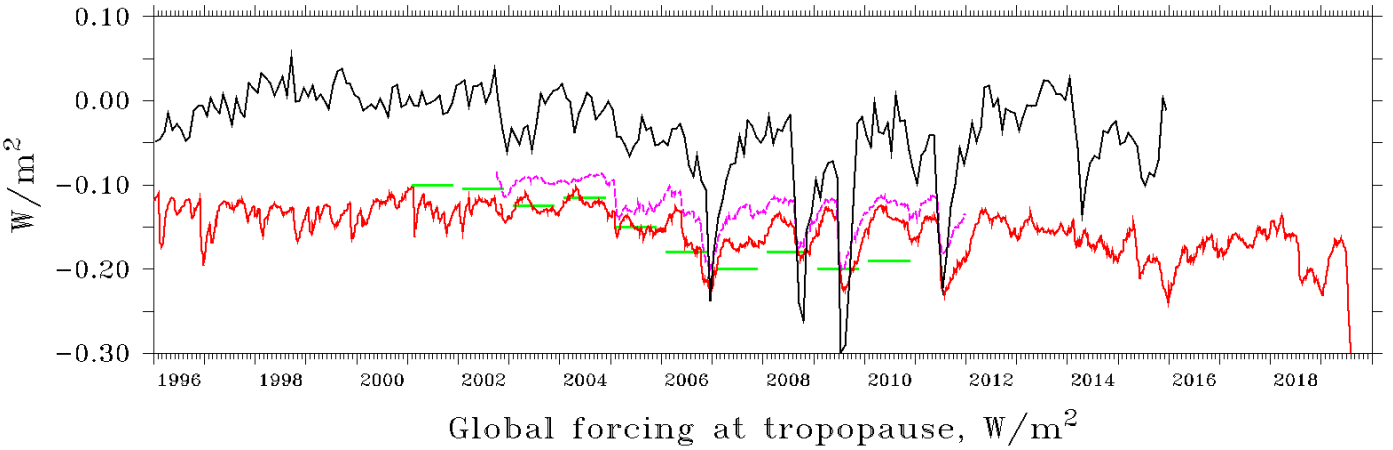
This is also one reason, why we can't set the integration boundary easily to the tropopause level.

It is helpful to have the SAOD results from Bruehl et al (2015) on Fig. 9. However, for a large portion of the 2002-2012 period of overlap, the difference between the results of Bruehl et al. and the present study seem very small, and it is not clear if there is any significant difference between them. They might both be within the uncertainty spread of the observations. There needs to be more clear evidence to support the conclusion that the new data set is an improvement upon prior emission estimates.

-> Response: Figures 9, 10 and 11b are presented with a logarithmic scale on the y-axis, so differences could seem very small. The differences are up to about 40%. See attached a linear version of the figure without Pinatubo.



Zoom in of Figure 9a. Stratospheric AOD at 550 nm wavelength: Tropical regions 20° S–20° N above 110 hPa are shown. Satellite observations from GOMOS (Bingen et al., 2017) are indicated by the green line and values derived from SAGE+CALIPSO (Santer et al., 2014) by the blue line. The red line shows the EMAC model simulations using the SO2 injections of Table 2, compared to the simulations of Brühl et al. (2015) (pink dashed line). The black line is from GloSSAC.



Zoom in of Figure 11a. Global radiative forcing by stratospheric aerosol. Estimated averages for solar forcing at the top of the atmosphere from satellite observations of annual averages derived from observations by Solomon et al. (2011) as green bars. The EMAC model simulations with instantaneous forcing at the tropopause (185 hPa, solar + IR) based on volcanic SO2 emissions are represented by the red line, compared to the simulations of Brühl et al. (2015) (pink line) and data from Schmidt et al. (2018) with volcanic effective radiative forcing (black line).

Specific comments:

All line numbers in reference to tracked changes document:

L10: “Directly” seems unfounded. If SO2 injections are derived from extinction measurements days to weeks after the injection, and derived using a model-based conversion factor and an empirical correction factor, is this “direct”?
-> Corrected: Removed "directly”

L16: the results of a simulation might be consistent with observations, not the simulations themselves.
-> Corrected: "results of the" simulations

L30: this sounds like a description of an externally mixed aerosol, which is incorrect.
-> Corrected: "internal liquid mixture"

L50-52: references needed for these statements.
-> Added citations: Aquila et al. (2012), Toohey et al. (2011)

L59: reference needed
-> Inserted: "(e.g. Kasatochi compared to Glantz et al 2014)", presentations at SSIRC and EGU meetings.

L66: This description of GloSSAC is so negatively phrased as to be potentially insulting to the authors of that work. Those authors have put years of work into filling the gaps, ensuring the best possible smooth transition between instruments, and addressing myriad other difficulties in the construction of such a long-term data set. It may not be perfect, and one can surely point this out, but I would encourage a more measured tone. Also, GloSSAC is primarily a climatology of aerosol extinction, and does not include SO2 at all. Therefore it is incorrect to say that GloSSAC excludes MIPAS data, since MIPAS does not provide measurements of extinction or the other optical properties included in GloSSAC.
-> Thank you for noticing this. Rephrased: "GLOSSAC (Thomason2018, Kovilakam2020), a time dependent aerosol climatology sometimes used for climate simulations, has a coarse temporal resolution and sometimes large uncertainties due to data gaps. It does not provide SO2 needed as input for chemistry climate models directly but only extinction."

L70: This makes no sense. It's true GloSSAC doesn't include SO2, so you don't need to explain why you don't use GloSSAC for SO2 injections. You do need to justify your work in relation to the other SO2 data sets that are available, including the work of Carn et al. (2016).
-> Response: In Appendix C a comparison of annual volcanic SO2 emissions from Mills et al. 2016 is added to the comparison between Carn et al 2016 and this study. To include that in detail in a table or a figure we lack data (since in Carn et al only a subset is provided).

L81: how is the location of the eruption a problem: the volcano location is known in almost all cases, is it not?
-> Response: Since simulations of point source emissions are very sensitive to the emission conditions, in some cases it may be more appropriate to implement the main plume of volcanic emissions in the model not directly at the volcano location, and instead choose other coordinates according to satellite observations. A case study for point source emissions is shown in Appendix C2.

L84: the logic here isn't clear: if tropospheric aerosols are removed by oxidation and rainout, how can they contribute to stratospheric aerosol layer at all, background or otherwise?
-> Added: most, "but not all" of the released SO2 is removed... and only a small fraction can reach the stratosphere by convection or large scale transport.

L98: It would be clearer for the reader to be as explicit as possible about the difference between which instruments are measuring SO2 and which extinction.
-> Response: Added "SO2 data" and "aerosol extinction data" to the single satellite instruments for clarification.

L171: the white areas in Fig 4 extend in latitude from the pole down to 0-30deg in both the NH and SH every year. The text here still appears to not explain this.
-> Sentences rephrased: "OSIRIS provides coverage from 82°S-82°N over the course of the year. Extinction is retrieved where the tangent point is illuminated, which is primarily in the summer hemisphere (see Figure 4)."

L215: SAD is a derived product included in GloSSAC, not "used in GloSSAC".
-> Corrected: "included"

L278: I still think it's worth being explicit that the IRF is calculated based on a difference of the two cases.
-> Corrected: Sentence rephrased: "Via multiple calls of the RAD submodel in one simulation, the instantaneous forcing is calculated online from the difference of fluxes for the cases with stratospheric aerosol only above 100 hPa and without any aerosol (Brühl et al 2012), additionally to the call with full aerosol used for the interaction with dynamics."

L280: The reader won't know what RAD_FUBRAD is if the sentence prior is removed.
-> Moved to Appendix A.

L319: "Observations and assumptions by Vernier (2016)" needs to be explained in much more detail.
-> Response: "like the decay of extinction by sulfate with time over 4 months."

L330: Are these limb or nadir measurement data sets" If the former, how does that affect consistency between those data and your results? If the latter, is your technique applied in full?
-> Response: nadir instruments (text modified)

L359: "integrating the vertical profiles" is fine, but to produce a single number for the SO2, you also need to integrate over the horizontal extent of the plume. This needs to be explained in detail.

-> Corrected: by integrating the 3-dimensional SO2 perturbations "over the boxes related to the volcano"

L374: I believe most use the term “Junge layer” to refer to the whole stratospheric sulfate aerosol layer, from the tropopause up, not just 25-29 km.
->Sentence rephrased: "Figure 6 shows the modeled vertical distribution of stratospheric SO2 in the Junge-aerosol layer with the local maximum of SO2 around 25 to 30 km altitude (Höpfner et al., 2013), typical mixing ratios of SO2 are about 0.03 ppbv."

L383: This might be true, but do you know for absolute sure? You mention there are less entries in the Smithsonian in the 1990s. Wouldn't this suggest it is at least possible there have been changes in eruption frequency?
-> Corrected: "is" -> "might be"

L385: It is true that at 17km, SO2 mixing ratios are generally larger in the tropics, but this may not be the case for all altitudes.
-> Added: in the “lower” stratosphere… See also next sentences.

L390: The impact of the Asian monsoon on Nabro aerosols has been addressed first by studies other than Clarisse et al., 2014.
-> Added citations: "Fairlie et al., 2014 and Bourassa et al., 2012”.

L393: Is the NASA SO2 database referred to here the same as that of Carn et al. (2016) which is referred to throughout the manuscript?
-> Added source: Online NASA SO2 database at: https://so2.gsfc.nasa.gov

L436: This might well be the case, but unless an experiment has been performed to test this, the statement should be softened (e.g., “…likely because…”).
-> Corrected: “…likely because…”

L437: The previous paragraph introduced 3 figures, it’s not clear which 2 are referred to here.
-> Corrected: "Figure 7 and Figure 8”

L448: Typically, stratospheric aerosol optical depth is calculated by integrating from the height of the tropopause up. Here, it is said that a lower bound is used that is 16 km in the tropics and 14 km in the midlatitudes. This is quite different definition, which needs to be justified. It is also quite unclear if this procedure is applied to both the model results and the observations.
-> Response: For practical reasons, the total stratospheric Aerosol Optical Depth (AOD) is obtained by the vertical integral of the aerosol extinction above an altitude of about 16 km (for mid-latitudes above about 14 km), since there is no information on the actual/variable tropopause altitude in the satellite observations to compare with.

L461: How do you know for sure? The GloSSAC data has been gap-filled. This may not be perfect, but it attempts to correct for the saturation. The difference between the model and GloSSAC then could be for another reason.
-> Response: “GLOSSAC tried to correct for that and has larger values.” In the data for July to September 1991 you still find remnants of the gaps due to saturation.

L468: What data product from AVHRR is included?
-> Response: “…the AOD from AVHRR (Advanced Very High Resolution Radiometer) by Long and Stowe 1994 at 630 nm is included.” The data were taken from a figure in that paper, there is no information on the data version.

L469: The AVHRR AOD product is at a different wavelength than the products shown on Fig 10. Any conclusion to be reached from the comparison may depend critically on the degree to which it would be “slightly less” or “slightly larger” when converted to the correct wavelength. Showing AVHRR, at the wrong wavelength on Fig 10, but not Fig 9 gives an impression of cherry-picking, showing data only when it appears to agree but not when it disagrees. If AVHRR is to be included, more effort should be put into making the comparison quantitative and fair.
-> Response: Text modified: “Consistent with the typical wavelength dependence, these values lie between the red curves for 550 nm (Fig. 9) and 750 nm (Fig. 10) at the peak after the Pinatubo eruption.” The curve is left out for practical reasons in Fig. 9 since it would cover other curves (see attached version).

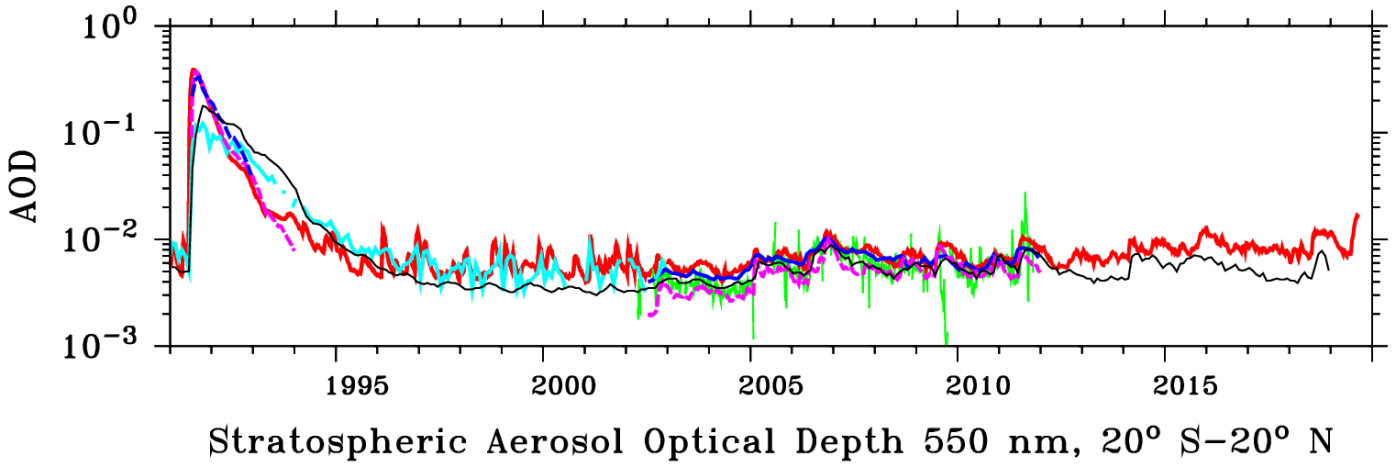


Figure 9a. Stratospheric AOD at 550 nm wavelength: Tropical regions 20° S–20° N above 110 hPa are shown. Satellite observations from SAGE II (Thomasonet al., 2008) are indicated by the light blue line, GOMOS (Bingen et al., 2017) by the green line and values derived from SAGE+CALIPSO (Santer et al., 2014) by the solid blue line. The red line shows the EMAC model simulations using the SO2 injections of Table 2, compared to the simulations of Brühl et al. (2015) (pink dashed line) and the black line in the upper panel is from GloSSAC. In this version we added the dashed blue line showing the AVHRR observations by Long and Stowe (1994) at 630 nm.

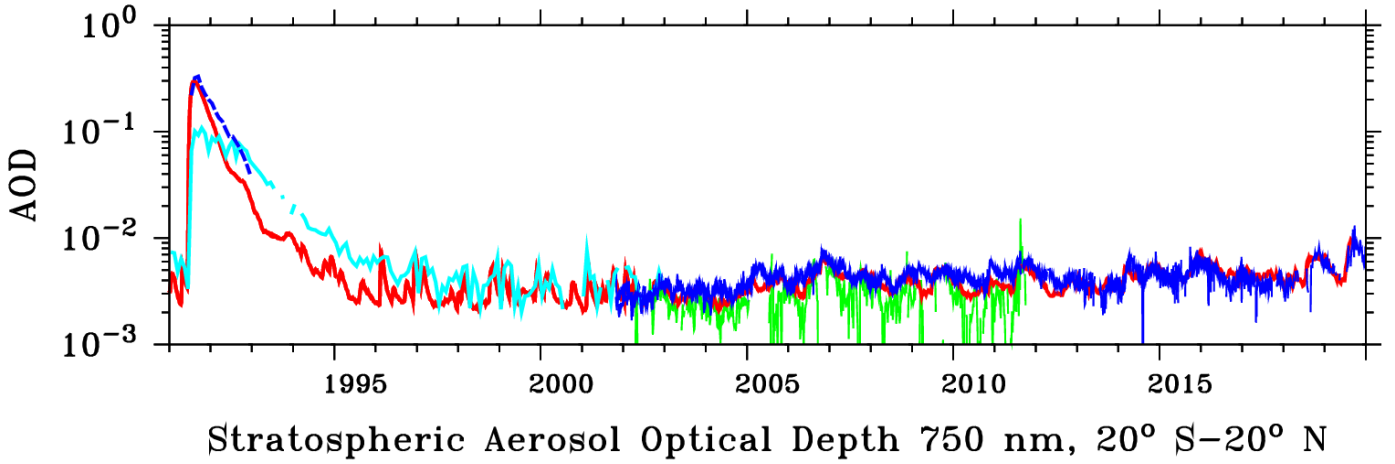


Figure 10a. Stratospheric AOD at 750 nm wavelength: Tropical regions 20° S–20° N above 110 hPa are shown. Satellite observations from OSIRIS (Rieger et al., 2019) are indicated by the blue line and GOMOS (Bingen et al., 2017) by the green line. For the Pinatubo period the blue dashed line shows the AVHRR observations by Long and Stowe (1994) at 630 nm. The light blue line shows the interpolation of SAGE data at 550 nm and 1025 nm wavelengths. The EMAC model simulations, using the SO2 injections of Table 2, are shown by the red line.

L471: this sentence seems to focus first on the Pinatubo eruption, but later on the impact of temporal resolution on the detection of individual eruptions. I doubt that temporal resolution has any impact on the magnitude of the peak AOD for Pinatubo. The argument needs to be clearer here.
-> Response: For better understanding, the sentence is split now. Of course, on the time scale of volcanic eruptions, which in some cases span only a few days, it makes a mathematical difference whether monthly means or multiple values with finer temporal resolution are considered; the same is held for zonal means. For a peak, the plotted maximum value is lower for a monthly average than for 5-day averages.

L503: “The large difference…” It’s not clear, but I think you refer here to the difference between your current simulations and the simulations of Schmidt et al. If so, then yes, the fact that the Schmidt et al. simulations used a smaller number of eruptions is *one possible reason* for the differences in RF in the periods between eruptions. But there are other possible reasons, including difference amounts of background aerosol injection, since a larger baseline in the Schmidt et al results would reduce the difference substantially. On the other hand, the model simulation from Bruehl et al 2018 also used a smaller eruption list than the present study, and the difference in RF between that simulation and the one of the current study look negligible in Fig 11. This would suggest that those smaller eruptions have a negligible impact on the RF.

-> Added: The large difference "of the simulations by Schmidt et al 2018" to our simulation…

L530: I'd suggest "infrared" rather than "thermal"
-> Response: The terminus "thermal heating rate", as well as "solar heating rate" is cited from Rieger et al 2020 (fig. 4). One may use "terrestrial" but "infrared" includes also solar infrared what is not meant here.

L540: "compute … through validation" makes no sense
-> Sentence rephrased: "… compute the volcano-induced radiative forcing using computed extinctions validated with satellite data."

L547: L2 needs to be defined
-> Corrected: "Level 2"

L571: This was not really demonstrated.
-> Response: Gaps in SAOD are not observed and background from OCs or other non-volcanic sources cannot explain the observations.

Fig B1, B2, B3: Over what time period are these values compiled? Are the plots to the left showing full zonal averages, or averages over the defined zonal extent of the plumes?
-> Response: Captions expanded:
"Figure B1. 2002 Reventador eruption: SO2 mixing ratio perturbation derived from MIPAS and the 3 extinction instruments. MIPAS data of 7 to 17 November 2002, OSIRIS and GOMOS data 11 to 22 November, SAGE II orbits of November and early December 2002. Zonal average …"
"Figure B2. 2010 Merapi eruption: SO2 mixing ratio perturbation derived from MIPAS, GOMOS OSIRIS and GOMOS with gap filling. MIPAS data 5 to 20 November2010, OSIRIS 14 to 25 November, GOMOS 25 November to 9 December (sparse data). Zonal average…"
"Figure B3. 2009 Sarychev and Mando Hararo eruptions: SO2 mixing ratio perturbation derived from MIPAS and OSIRIS., arrangement of panels because event is split at 24°N as for integrated values in Table 2. MIPAS data 18 June to 18 July 2009, OSIRIS data 17 July to 3 August."

Fig B3: These plots seem to clearly show a continuous single plume. It may be the combined impact of two eruptions, but to split the plume at 24N seems extremely arbitrary.
-> Response: The event has been split at 24°N as for integrated values in Tab.2. (see above)

Table C1: How is explosive defined? Column "in %" needs to be better defined/described.
-> Response: Column "in %" is the fraction of explosive emissions from total emissions, now removed to get space for comparison to Mills et al. 2016. The term "explosive" SO2 emissions is used by Carn et al. 2016 [Table 3] for "explosive eruptions only (i.e., excluding effusive eruptions)". In the caption we have added contributions from outgassing (Diehl et al, 2012).

Report #2

Referee #2: Thomas Aubry, ta460@cam.ac.uk

First, I thank the authors for their efforts in addressing my comments and for significant improvements to the paper. However, one of my main suggestion has not been addressed. The ultimate goal of this paper is to take advantage of a range of observational capabilities to conduct simulations in which volcanic SO2 is injected as a spatially-resolved SO2 cloud instead of the traditionnal point source approach. I think this is an interesting and valuable idea, but I find that the paper is undermined by the fact that no equivalent simulations with the same model and with point source injections were run. This makes it difficult to understand and discuss the advantages and limitations of the methods proposed by the authors. This is the last major comment I have, and I otherwise find the manuscript acceptable for publication.

-> Response: In Appendix C2 we added a chapter with sensitivity simulations for comparison of different case studies for volcanic point source injections (about 5 pages including figures). This is now also mentioned in the abstract, introduction, section 5 and the conclusions. Appendix A includes an additional module for that.