Dear Reviewer,

Thank you very much for your careful review and helpful comments. We have addressed each point below. Reviewer comments are in **black**, and our responses are in **blue**.

Best wishes, Christine Pohl, Ulrike Niemeier and Co-Authors.

Overall review:

This manuscript presents a comparison study between various global circulation models and satellite measurements made. This study focuses on comparing modeling and satellite observation data for stratospheric aerosol emitted from several volcanic eruptions. There is clear scientific merit presented in this study but the authors do not clearly motivate their claims or justify their stated significance with results. They do not present an inter-comparison with satellite observations, but those comparisons are not quantified and only discussed in passing. Additionally, the satellite results are assumed to be uncertain without justification. In no part of the comparison is there a discussion or quantification of data agreement. There needs to be a quantification of how the data are comparing. The organization of the paper is clear, but there are various typos in the figure cations and there is not clear demonstration as to how they are judging data agreement. Currently their is no quantifiable justification for their abstract statement. I recommend **major revision**. Below are general comments and suggested changes. In addition to general comments and suggestions, I have also made specific comments and suggestions within the attached manuscript PDF file:

We apologize for the typos we made. Spelling errors were corrected (e.g., L. 120/121), wording errors in both Figure captions (Fig. 2, 4) and definition inconsistencies in the Figure legend (Fig. 8) are eliminated.

#Abstract:

L.2: This seems like a long way of saying global circulation model (GCM). Suggest rewording.

Changed to: circulation model simulations with aerosol–microphysical processes.

L7: This seems like and odd statement for a abstract. Suggest removing and replacing this text with some significant results/findings.

We have removed the sentence.

L10: “Strong Consistency”: This seems vague. Suggest showing metrics for representing consistency would be helpful to illustrate your claims.

We thank for your comment and agree. We have revised the abstract.

L15ff: What were the improvements in the important comparisons? This is unclear. Suggest adding information as to what was compared and what were the significant results. This seems

like a completely non-physical result. Why would this result be important? This also seems unclear. What improvements are you suggesting and why?

We thank for your comment and agree. We have revised the abstract, accordingly.

#Introduction:

L37: Is this the same thing?

No, ESMs and GCMs are different, as the latter simulates atmospheric processes only. We changed the text to:

Model simulations enhance our understanding of the role of volcanic sulfate aerosols in the climate system.

Earth system model (ESM) include e.g. atmosphere, land, and ocean models, necessary to simulate the climate impact of volcanic aerosols. To reduce the computational amount, they often rely on prescribed aerosol optical properties....

L41: What module? Would any aerosol module work or just one? Suggest changing "the..." to "an aerosol...:

Changed to:

The module simulates the evolution of the aerosols, such as particle formation and growth, as well as their radiative interactions.

L42: This is confusing to me. Is there more than one mechanism? As written the heating due to particle formation is the only mechanism you mention can be studied. Suggest adding "such as" and rewriting the following text accordingly.

We agree here and changed the text to:

This allows the study of stratospheric aerosol feedback mechanisms, such as the relationship between particle formation and the dynamic impacts resulting from the heating by absorbing radiation.

L73: This is unclear. Suggest adding some quantifiable values on the size growth relative to dry.

… which can significantly enlarge N by a factor of 10 to 1000, depending on the size of the volcanic eruption.

L. 78: model results support the evaluation of … Changed to:

model results provide confidence in the SCIAMACHY-retrieved aerosol properties

L.79: We demonstrate the robustness of the SCIAMACHY retrievals. Changed to:

We demonstrate the reliability of the SCIAMACHY retrievals

#Methods:

L119: Define T, SO2 concentration range.

We provide T, SO2 concentration ranges:

The parameterization is valid for low stratospheric temperatures, up to -80°C, and high SO_2 concentrations in a volcanic cloud, which can be increased by several magnitudes above the background conditions.

Table 1: This reference is never defined. Suggest using a proper data or publication citation.

A specific reference to volcano.si.edu is given.

L. 124: Why use this value? Suggest adding clarity from the common literature values.

These values follow the standard set up of the aerosol module HAM (Stier et al., 2005). We cannot give a reasoning for the value they had chosen.

L.125: How much stronger were those? Suggest adding some context from the paper here.

We have revised the sentence:

… similar to previous studies of stronger volcanic eruptions like the eruption of Mt Pinatubo in 1991.

#Results:

L149: What is meant by remarkable agreement? Can you discuss some statistics like R2, chi2, MSR, etc.?

We have revised the results section (Sect. 3) to include important statistical measures and added the results to the Figures 1-4. We focused on symmetric measures of agreement (eg. Pearson correlation coefficient, RMSE). We did not use non-symmetric measures (eg. Chi_squared), as these require one dataset to be defined as the reference, whereas in our comparison, satellite observations and model simulations are treated as equivalent. We also note that a statistical assessment of the SCIAMACHY-ECHAM comparison is challenging because the observed and simulated aerosol plumes are spatially shifted. Although these shifts are generally small, they

vary with time. This, combined with a – for a statistical analysis relatively – small data amount means that not all statistical measures can be meaningfully applied to every comparison presented in the manuscript.

#Discussion:

L202: What is meant by best agreement? Can you discuss some statistics like R2, chi2, MSR, etc.?

Providing statistics for Figure 5 and 6 is challenging due to the sparse vertical resolution of the measurement data (3.3 km) and the slight spatio-temporal mismatch between the simulated and measured aerosol plumes. We therefore focus on percentage deviations:

The comparison of Ext demonstrates that the SCIAMACHY v2.0 aerosol PSD retrieval provides the best agreement with SAGE II among the evaluated datasets. Relative differences within the plume with respect to SAGE~II remain within about +/-20 % during the first observation period and reach +/-40 % during the later periods. Modifying the number density profile in the SCIAMACHY retrieval only affects the magnitude of Ext, while its spatial distribution remains unchanged. This increases the absolute relative difference by approximately 20 percentage points. Among the three SCIAMACHY datasets, direct Ext retrieval (SCIA Ext, Fig. 5 m-o) yields the lowest values, with relative differences to SAGE II ranging from approximately -40 to 20 % during the first observation period and increasing to -60 to -15 % during the later periods.

… Correspondingly, relative differences with respect to SAGE II are between -80 and 80 % during the first observation period and decrease to -50 and 25 % in the third observation period.

L.211: This is a pretty big statement that should be justified with estimations on the uncertainty.

We have revised the sentence:

As this approach involves two retrievals, their respective uncertainties may contribute to a higher overall uncertainty into the retrieved product.

#Conclusion:

L.313: This needs to be quantified. This is not effectively done and should be based on the author's own claims in the introduction.

We have revised the first part of the Conclusion:

This study compares aerosol optical properties, Ext and reff , retrieved from SCIAMACHY with ECHAM simulations following the Manam (2004/05) and Sarychev (2009) eruptions. The comparison serves two complementary aims: first, to provide confidence in the SCIAMACHY-retrieved aerosol properties despite the assumption of a fixed background N profile, and second, to assess the performance of ECHAM and improve understanding of the processes governing particle size and plume transport.

Good agreement is found for Ext and reff after the Sarychev eruption, with a Pearson correlation coefficient of 0.89 for Ext. The simulated and observed plumes show similar horizontal and vertical extents, while the evolution of reff is consistent between the two datasets, including both particle growth and shrinkage. At 18.4 km, the mean Ext difference remains below 20 % during the first months after the eruption, gradually increasing to about 50 % by the end of 2009. The retrieved and simulated reff values agree within about ±10 %, decreasing to about ±5 % over time. For Manam, the agreement is weaker. The Pearson correlation coefficient between simulated and observed Ext is 0.71, with mean differences of 20 - 40% at 18.4 km and increasing to 20-50% at 21.7 km. The simulated plume remains about 10° farther south than the observed one, contributing to the larger Ext differences. Differences in r_eff reach up to 20 %.

The consistency between SCIAMACHY v2.0 PSD retrievals and ECHAM simulations, particularly following the Sarychev eruption, provides confidence in the SCIAMACHY-obtained Ext and reff, despite the assumption of a fixed N profile. Comparison with SAGE II data further shows that the SCIAMACHY v2.0 PSD retrieval yields Ext values in better agreement with SAGE II than the direct SCIAMACHY Ext retrieval, indicating that the v2.0 PSD approach provides more realistic Ext. The provision of aerosol size information is an additional advantage of the SCIAMACHY v2.0 PSD approach. Modifications in the assumed N profile slightly affect the magnitude of Ext and reff, but have little influence on the vertical profile shape. Given the limited availability of remote-sensing observations of aerosol size, SCIAMACHY v2.0 PSD therefore provides a valuable data resource for studying stratospheric aerosol evolution.

L.323: “A model parameter study highlights the importance of using realistic injection profiles and background particle sizes in ECHAM for improving model–observation agreement.” This seems to not be supported by the results discussed here. Some statical quantification is needed. Are the observations from SAGEII? This is barely discussed and the authors even state the agreement is poor.

We revised the sentence, also according to a comment of reviewer 1:

The ECHAM parameter study demonstrates that the injection profile and background particle sizes strongly influence the simulated aerosol evolution and are important for achieving better agreement with SCIAMACHY.

L.335: This does not seem justified. There are no statistics supporting this.

Indeed, the sentence was misleading. We have revised it. Statistics are provided at the beginning of the Conclusions:

It also provides confidence in the SCIAMACHY v2.0 PSD retrievals despite the assumption of a fixed N profile.

L337: This is correct, but what has your work done towards this end? The majority of the conclusion discusses setup and motivation but does not discuss the significance of the results and findings of the discussion.

We thank the reviewer for this comment. We have revised the last paragraph of the conclusion, accordingly:

The parameter study demonstrates the complexity of realistically simulating the evolution of volcanic aerosol plumes and the need to better understand and accurately parameterize the underlying aerosol-microphysical processes. This is particularly relevant for stratospheric aerosol intervention (SAI), where the same models are used to predict the evolution and radiative effects of artificially injected sulfur, but cannot be tested against a real intervention.

Overall, comparison studies like the current one are essential for building confidence in both satellite observations and model results.

General comments:

The following aspects are those considered important by AMT for review:

3. Are substantial conclusions reached? No quantifiable conclusions are made.

5. Are the results sufficient to support the interpretations and conclusions? The do not quantify their results with and statistics or clarity. Most of their statements are vague and unclear.

We thank the reviewer for both valuable comments and acknowledge these concerns. We have revised the manuscript to address them by including statistical measures and quantitative results in the comparison section. Additionally, statistical quantification has been added to the Abstract and Conclusion, and vague statements have been revised to improve clarity and precision.

6. Is the description of experiments and calculations sufficiently complete and precise to allow their reproduction by fellow scientists (traceability of results)? Not really. There would be no way for someone to replicate their results.

We thank the reviewer for raising this important concern. SCIAMACHY v2.0 aerosol PSD retrieval data is available at https://www.iup.uni-bremen.de/DataRequest/. In addition, the ECHAM simulations will be made available at https://www.wdc-climate.de/ui/entry?acronym=DKRZ_LTA_550_ds00008. (Data will be available at this link soon.) This now enables the reproduction of our results.

9. Does the abstract provide a concise and complete summary? The abstract claims are concise but not represented well in the manuscript's contents. Suggest clearly quantifying the claimed results and altering the abstract to align with the manuscript.

See answers above.

10. Is the overall presentation well structured and clear? In general, the work is structured appropriately but the comparisons seems unclear.

See answers above.

11. Is the language fluent and precise? There are some errors and typos, but the language is reasonable but not precise. Plenty statements of good agreement but no quantification is given to justify the language.

We have replaced statements of "good agreement" to quantitative statements.

13. Should any parts of the paper (text, formulae, figures, tables) be clarified, reduced, combined, or eliminated? Most of the text needs to be changed to include specific discussion of quantifying how different metrics agree.

Done, see answers above.