We would like to thank the anonymous referee for taking the time to read and provide comments on our paper. We believe the changes made in response to their comments have strengthened the manuscript.

See below for our response (in red) to their comments (in black).

Response to anonymous referee

For two months of the Holuhraun fissure eruption in 2014 and 2015, this study presents an analysis of the sulphur and sulphate dispersion over Europe as well as an evaluation of the sulphate production. For this, dispersion simulations of multiple models are utilised, compared, and evaluated against IASI SO2 retrievals and surface concentration measurement of European monitoring stations. These comparisons are fairly conducted and discussed. However, to a large extend the authors seem to focus on model comparison and evaluation. Consequently, the manuscript could partly be more adequate to be published in GMD rather than ACP. However, the investigation on SO4 formation fits well with the purpose of ACP and should be emphasised more in the abstract. Furthermore, there are multiple aspects in the model intercomparison, the trajectory-based source assignment, and the SO2-to-SO4 conversion that need further elaboration before the manuscript may be published.

Major comments:

A multi-model intercomparison necessarily needs to consider differences in the models' characteristics as model physics/dynamics/chemistry and the model setup as parameterisations and model resolution to allow for a fair comparison and meaningful discussion. Therefore, I suggest adding the advection schemes used in Table 2. How can the OsloCTM3 not have a chemistry/aerosol module as a CTM? Furthermore, the vertical model level distribution is essential with respect to the volcanic emission plume. Please include the model layer thickness of the lowest model level and the number of model layers between 0 km and 3 km height in section 2.3 and consider this in the discussions! How do the different models differentiate with respect to vertical layering, vertical distribution of emissions etc.? Further, it is very hard to assess the performance of the different models individually, as the evaluation and the figures/tables often include just a selection of the models. It seems to be not a full and fair comparison between all models. Figure 2 does not include the CTM due to missing required diagnostics (lines 290-291). What does this mean? Why is it then being listed in table 3? Figure 3 just evaluates the performance of 3 models. Why are the model outputs not designed in the way that they are comparable? Please consider producing comparable model output to fully discuss all model performances.

Thank you for your suggestions on how to improve the model inter-comparison part of our study.

Firstly, you are correct in that the model vertical resolutions need to be considered more. Subsequently, details on the model vertical resolution have been added to Table 2 and considered during the discussions. However, we have not included the advection schemes. As the models are nudged to ERA-Interim reanalyses data, the dispersion is already constrained which was done, partly, to mitigate the transport errors. Hence, we feel added details on advection schemes is superfluous.

Secondly, you are right that OsloCTM3 has a chemistry scheme, and its omission is an error that has been corrected.

Moreover, all figures inter-comparing model output now contain all models where possible.

Finally, the SO_2 columns loads and mass burdens for OsloCTM3 have now been included and the discussion amended accordingly. Unfortunately, from the diagnostics made available in the submission for this experiment, it is only possible to derive monthly SO_2 plume height estimates for MIROC6.1-SPRINTARS and ECHAM6.3-HAM2.3-P3, whilst no estimate is possible for OsloCTM3. Whilst we agree with the reviewer that having all the diagnostics requested for the experiment would be ideal, the main aim of the experiment is the investigation of the aerosol-cloud interactions (ACIs) in Part 2. As a significant number of diagnostics were requested and the submissions made "pro-bono", some modelling centres focused their resources on the diagnostics required for the ACI study leaving some diagnostics absent for Part 1. We agree this limits the plume height inter-comparison, yet still believe the discussions on the output available to the experiment warrants inclusion.

The setup of the backward trajectories simulated with HYSPLIT needs clarification (lines 174-178). It remains unclear how the ensemble trajectories are designed. Are there 27 ensemble members defined for each station and each hour? If so, how are the ensemble members at the individual stations perturbed? Or do you have one ensemble member per hour for each station? However, 27 stations do not agree with table 1. Table 1 lists the starting heights for the trajectories at each station. How are these heights defined? Is there a basis for these heights as the station height above the surface height of the corresponding GCM's grid cell? How do you ensure that the trajectories allow for a fair comparison with ground-based observations? Further, are 2 % of the trajectories passing through the 3D bounding box (line 184) significant enough to connect these to the Holuhraun event?

Thank you for raising your concerns regarding the trajectory analysis. We hope the following changes reassure you that our methodology is sound.

Firstly, a more detailed description on the trajectory analysis has been provided. This includes a clarification on how the 27 ensemble members differ due to small perturbations in the input meteorology data (i.e. offset by a fixed grid factor, a maximum of 1.0° of latitude/longitude in the horizontal and 0.01 sigma units in the vertical, and so all possible offsets result in the 27 members).

Secondly, the trajectory starting heights for each station used offsets to the model surface to address differences between the orography in the driving meteorological data and reality.

Thirdly, we agree that the 2% threshold was too small to be confident in attributing a pollution event to Holuhraun emissions. The threshold has since been increased to 25%.

The assessment of the transport time (lines 185-192) is not fully clear and needs some rephrasing. I would have expected a circular influence region around the eruption site instead of a squared bounded domain. Please also elaborate on the definition of the idealised trajectory points in Fig. 1. Do these relate to full hours? And how many black circles are attributed to which trajectory?

We agree that a circular influence region is more suitable and have made changes to our trajectory framework to include such. In our revised framework, the idealised trajectory points in Fig. 1 are no longer relevant and have been removed.

Regarding the IASI SO2 retrieval, it remains unclear what “the SO2 detection is positive” means (lines 112-114). Is 0.49 DU a detection limit? Or is it an individually defined value to discriminate volcanic SO2 from other SO2 (which is the climatology)? Please clarify this. Furthermore, in line 118, you are referring to meteorological temperature profiles for the height conversion. Where do these profiles come from? Are they standard profiles or from meteorological model analyses? To better understand the uncertainty of the IASI retrievals, it would be desirable to have a short summary of the different components contributing to the retrieval error. For example, do the uncertainties of assuming Gaussian profiles and the uncertainties of the temperature profiles contribute to the retrieval error? With respect to the discussion in line 249, please justify why the central height of a Gaussian SO2 vertical profile can be an estimate for the injection height. There are enormous amounts of mass being distributed above the central height of a Gaussian profile.

Thank you for seeking additional clarity on the IASI SO_2 retrievals.

The 0.49 DU is a threshold set by Carboni et al., 2019 within their IASI retrieval algorithm and is chosen specifically for the Holuhraun eruption. The threshold is used in the detection scheme which is a linear retrieval with the SO_2 column load as the free parameter. The linear retrieval assumes the background SO_2 concentration follows a Gaussian distribution. The threshold is set substantially higher than the standard deviation of this distribution meaning a positive result is exceedingly likely to be significantly different to the background. We have improved the clarity of this section within the manuscript. Further details on the detection scheme can be found in Walker et al. 2011, 2012.

In terms of the height conversion, this is performed within the algorithm in Carboni et al., 2019 and is not a post processing step by us. They use atmospheric profiles from European Centre for Medium-Range Weather Forecasts (ECMWF) meteorological data.

We have added extra details on the IASI error estimates and provided references to seek further information. Although considered important, the reason for using the Carboni et al., 2019 product was to seek information on the plume location rather than getting absolute concentrations right. Hence, an extensive summary on the individual components contributing to the retrieval error is superfluous to our goal.

We have removed our comment on the injection height estimate after further consideration.

The core element of the manuscript is the investigation of the SO2-to-SO4 reactions. The use of the biexponential fit and the division into gas-phase and aqueous-phase pathways seems promising on a first sight. Are there any references, where you base this method on? When exploring Figure 6a, the monoexponential and biexponential fits appear very similar and it is hard to justify why the biexponential fit performs better. How can you make sure that the two reaction pathways can directly be mapped within a biexponential fit? The scattering of the data point remains widely spread while the exponents derived from the biexponential fit are fairly close. Please re-evaluate this analysis and provide more evidence for the pathway assumption.

We thank the reviewer for raising their concerns with the biexponential fitting. A similar comment was made from another reviewer. In hindsight, we acknowledge that attempting to separate the in-plume SO_2 into its gaseous and aqueous pathways was too much of stretch for our dataset. We now only fit to an exponential with a single decay constant. The derived gaseous-phase and aqueous-phase oxidation rate constants are now replaced with a single value generalising them. However, we

believe that our efforts to explore the complexity of volcanic SO_2 oxidation is still worth mentioning and a suggested direction for future works has been included.

Minor comments:

Please review all citations (e.g., in line 50 Aas et al., 2015 is cited, but does not exist in the reference list).

Corrected, thank you.

Please check punctuation. Extra commas would increase readability. And e.g., line 209 misses a "." after "respectively".

Corrected, thank you.

Regarding the tables' captions, these are typically written above the tables. Please revise.

Corrected, thank you.

Line 36: "ugm-3" must probably be µgm-3

Corrected, thank you.

Line 55: delete doubled "aerosol"

Corrected, thank you.

Line 111: SO2 column load and plume height "are" derived…

Corrected, thank you.

In line 138, the calculation of monthly surface mass concentration climatology is not fully clear. What time span is used here? Is the full temporal coverage mentioned in the text corresponding variable for the different stations and corresponds to the column "Temporal coverage" in table 1?

Yes, you are correct. We have amended this explanation within the manuscript to provide added clarity.

Table 1: I just count 22 EMEP stations being listed in the table.

We find 25. Perhaps part of the table has not rendered probably during upload as it spans multiple pages.

Figure 2: Why mentioning the 21 UTC sampling in the caption, if the figure shows simulation results in the morning?

Corrected, thank you.

Lines 264-265: Here, sharp peaks and troughs are mentioned but probably only the troughs are discussed. This is confusing. Please also check the dates listed here. These are not well recognisable in Fig. 3c.

This statement has been removed as is no longer required and the dates have been checked.

Figure 4: Please define "pollution event". What is the timeframe of high sulphur concentrations for such an event? And can events occur multiple times a day?

"Pollution event" is defined in Sect 2.2. and we have added a comment to the figure caption to clarify this. No, only one event per day is considered here.

In lines 337, 387, and 401, there are 22 EMEP stations mentioned. However, the explanation before states 20 stations. Please check!

The change in the number of stations is because we are discussing different totals in these sections. To clarify, this study uses data from 25 EMEP stations. Of the 25 EMEP stations, 22 experience at least one pollution event, regardless whether it is Holuhraun attributed or not, during September and October 2014. Then, of the 22 EMEP stations experiencing a pollution event/s, 19 experience at least one event that has been attributed to Holuhraun.

Or put simply, the station counts change as follows:

- 25 -> 22 as 3 EMEP stations did not experience a pollution event of *any* origin
- 22 -> 19 as 3 EMEP stations did not experience a pollution event of Holuhraun origin

We have added further clarity within the manuscript to address this confusion.

Line 371: Meaning of "… as the models are essentially trying to correctly capture the behaviour of two pollutants as opposed to one" is unclear. Please rephrase.

Our statement was flawed and has been removed.

Line 425: Please add SO2-to-SO4 again before "ratio of 31+-4".

Corrected, thank you.

Table 5: Should ECHAM6.w-HAM2.3 have a footnote indicated by the "*"? If yes, where is the explanation?

This was used to identify that this model's output could not be fitted successfully to a biexponential fit. Since we no longer do this fitting, the "*" has been dropped.

Line 462: A comparison against IASI SO2 retrievals "shows" that…

Corrected, thank you.

Lines 464-465: Please be more precise here. What is an underestimation of a distribution?

Corrected, thank you.

Line 473: "whilst considering everything else equal" Is this really the case? What is about the different resolutions, different chemical mechanisms, different transport schemes? Please extend this discussion.

We have reworded our comment for clarity. The intent was to state the importance of understanding the volcanic perturbation to the region before exploring the impacts this perturbation has on aerosol-cloud interactions, rather than to discuss the model differences.

Carboni, E., Mather, T. A., Schmidt, A., Grainger, R. G., Pfeffer, M. A., Ialongo, I., and Theys, N.: Satellite-derived sulfur dioxide (SO2) emissions from the 2014–2015 Holuhraun eruption (Iceland), Atmos. Chem. Phys., 19, 4851–4862, doi:10.5194/acp-19-4851-2019, 2019.

Walker, J. C., Dudhia, A., and Carboni, E.: An effective method for the detection of trace species demonstrated using the MetOp Infrared Atmospheric Sounding Interferometer, Atmos. Meas. Tech., 4, 1567–1580, doi:10.5194/amt-4-1567-2011, 2011.

Walker, J. C., Carboni, E., Dudhia, A., and Grainger, R. G.: Improved detection of sulphur dioxide in volcanic plumes using satellite-based hyperspectral infrared measurements: Application to the Eyjafjallajökull 2010 eruption, J. Geophys. Res., 117, D00U16, doi:10.1029/2011JD016810, 2012.