We thank the reviewer for her/his positive comments on our work and all the suggestions which improve our work. Please find our replies insert in red.

Title: The IASI retrievals are also uncertain and quite different between the two products. Also the CALIPSO aerosol heights may or may not have the same vertical distribution as SO_2. Given the uncertainties in these independent datasets, one may argue that the study here is more of a "comparison" rather than a "validation".

Following a similar suggestion from the other reviewer, we have changed the title to: Volcanic SO_2 Layer Height by TROPOMI/S5P; evaluation against IASI/MetOp and CALIOP/CALIPSO observations.

Lines 151-154: With the different overpass times between IASI and TROPOMI, why not use trajectory model to match measurements between different sensors?

This is a very good idea. We are currently working on implementing a trajectory/dispersion model in order to track and forecast the volcanic plume. This could be also used to correct for different overpass times. We are however still in the process of developing this and cannot simply apply this in the given paper. This is nevertheless foreseen in the near future.

Figure 2: Are the integrated profiles based on the same grid cells (i.e., for grid cells that have both valid TROPOMI and IASI height retrievals)? Or is the mass difference between IASI/AOPP and TROPOMI due to different pixels being integrated? What could be the reason for different SO_2 mass estimates between TROPOMI and IASI/AOPP? Please clarify.

The IASI AOPP algorithm quality control rejects pixels within the core part of the plume, due to the poor fit between the measured and modelled spectra. The SO_2 spectral lines chosen by the IASI AOPP algorithm get saturated by the large SO_2 amounts and the retrieval fails to pass the quality control. This is a known fact to the IASI AOPP algorithm scientists and a different algorithm set-up to amend this issue is currently work-in-progress. As a result, when all pixels are excluded in IASI AOPP for a single grid box, the grid box is excluded from the presented comparison and, due to the particularities of the algorithm discussed above, very high concentration pixels are excluded by the IASI AOPP quality control. This in turn lowers the IASI AOPP SO_2 mass estimate, which is not the case for the IASI ULB/LATMOS dataset.

Figures 2 and 3: The distribution of TROPOMI retrievals is more spread-out - do we know why?

The main reasons for a broader distribution of TROPOMI retrievals wrt IASI is quite likely the different wavelength range sensing the plume as well as the use of a completely different retrieval approach, i.e. NN vs optimal estimation. In the NN L2 regularization is applied such that it generalizes better, especially since simulated reflectance spectra are used for the training of the NN. This has some impact on the spread of the LH retrievals. Secondly the IASI LH retrievals are sensitive to a different altitude range in the IR than that of the UV wavelength range used in the S5P LH retrievals, which in turn has also influence on the LH distribution.

Lines 324-325: The comparison sample is dominated by Raikoke - can the authors elaborate how this affects the comparison (for example, correlation coefficient)?

It is indeed inevitable that the validation is biased towards Raikoke, since it was the strongest and most long-lasting eruption during the time period of our study. By removing the two other days of eruptions from both comparisons of Figure 4, the statistics do not alter significantly. For IASI AOPP, the slope is 0.90 [0.91], y-intercept of 1.40 [0.90] and correlation coefficient of 0.631 [0.66]. For IASI ULB LATMOS, slope of 1.10 [0.98], y-intercept of -0.45 [0.77] and correlation coefficient of 0.663 [0.72]. In parenthesis I provide the statistics of Figure 4.

Figure 6: Please specify the thresholds used to filter out CALIPSO weighted extinction height data (clouds?)

To avoid cloud contamination of aerosol retrievals, cloud signatures must be identified and removed. Prior to analysis, advanced QA procedures are performed on the L2_05kmAPro product to remove highly uncertain aerosol extinction data values. This QA scheme is similar to that employed in Campbell et al. (2012) and Winker et al. (2013) and involves several parameters included in the L2_05kmAPro product:

- (1) Extinction_QC_532 (r) is equal to 0, 1, 2, 16 or 18,
- (2) $-20 \geq$ CAD_Score(r) ≥ -100,
- (3) Extinction_Coefficient_Uncertainty_532(r) ≤ 10 km^{-1}
- (4) Extinction_Coefficient_532(≥ 0 and ≤ 1.25 km^{-1}),

Further details of each QA parameter are documented in the CALIPSO Data Users Guide (http://www-calipso.larc.nasa.gov/resources/calipso_users_guide/).

Also Figure 6: Are stripes in TROPOMI SO_2 heights due to retrievals or gridding?

Thanks for the comment. The "stripes" in the TROPOMI SO2LH map in Figure 6 are due to a simple visualization of the TROPOMI pixels via Python. Each color grid point represents the center of TROPOMI pixel so there are "white" areas left between pixels. It is not related to any gridding process.

Figure 7: Can the authors use different colors for the data points in the right panel based on SO_2amount?

We have updated Figure 7 according to the reviewer's suggestion. The new figure is color-coded according to the corresponding TROPOMI SO2 VCDs while range between 21 and ~120 D.U. for that day.

Figure 8: Can the authors comment on the low correlation between CALIPSO and TROPOMI? Is the correlation coefficient a function of time since the main eruption? Based on Figure 6 and Figure 8, can we draw the conclusion that individual TROPOMI retrievals are not so well-correlated with CALIPSO measurements?

What is immediately apparent from our analysis is that the CALIOP ALHext is higher than TROPOMI SO2LH almost consistently for most of the cases As volcanic aerosol layers evolve and disperse into the atmosphere, their optical and microphysical properties are expected to change with time. We see a systematic increase in the average daily heights in CALIOP measurements, which is not the case with the TROPOMI observations (Table 4). It should also be noted that the number of collocations vary significantly day-by-day. We cannot however argue that there is a “clear” dependence over time for the correlation between the compared datasets.

The obvious reason is, as discussed in this work but also other studies that compare CALIPSO to Uv/Vis instruments volcanic observations, the SO_2 and ash/aerosol plumes are not necessarily collocated, as gas and ash can separate in volcanic ash, especially in the days following the eruption. After they separate, it is up to the prevailing winds in the region to either separate them further, or bring them at the same altitude once more. Ageing of the ash particles also plays a significant role, further complicating the issue.

Another reason could be due to CALIOP possibly underestimating the aerosol layer thickness due to strong attenuation of the lidar signal at the top of the layer (Rajapakshe et al., 2017), whereas the TROPOMI SO2LH product does not suffer from such attenuation. It is well known that the CALIOP based layer detection often misses the lowest boundary of a thick aerosol layer, thereby biasing the bottom of the aerosol layer high. We have to note here that the CALIOP is able to measure only ash and aerosol absorption profiles.

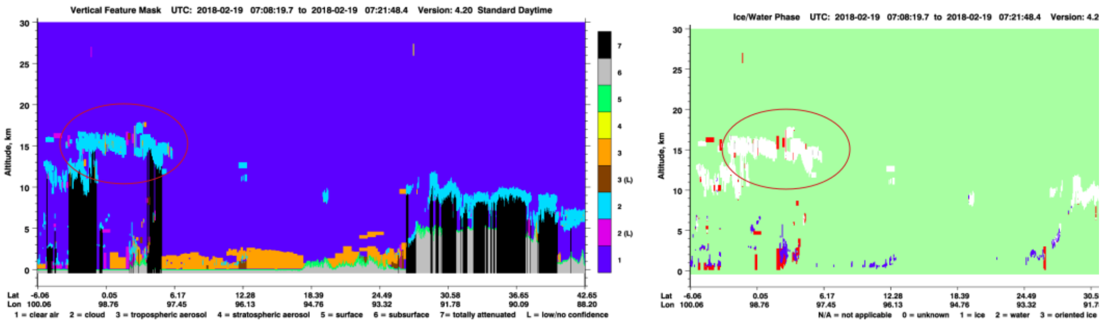
Section 4.2.2 focuses on Sinabung but the discussion (lines 397-405) appears to indicate that the eruption is not an ideal case for validation/comparison using CALIPSO?

From our extensive analysis of this volcanic eruption during the S5P+I: SO_2 LH project, we have concluded that this case is not an “ideal” case for a direct comparison of the TROPOMI-CALIPSO datasets. We selected to present this case study for Sinabung in the paper to demonstrate a situation where is characterized by complexity. This case of mixing between ash and clouds over a volcanic eruption renders the retrieval of the ash plume altitude by the lidar algorithm very difficult, since it cannot separate clouds from aerosols, especially when the aerosol amount is low. Although the eruption was spatiotemporally small an excellent overpass was found against the CALIPSO instrument. Performing the validation of the Sinabung eruption (12 January 2018) we used Total attenuated backscatter (TAB) (and show in the manuscript) to provide a qualitative analysis for the SO_2 LH retrievals. We summarize the main points arising from this case study:

- We expect that, in case of sufficiently dense ash, the cloud height data products provide accurate volcanic ash cloud heights. On the other hand, In case of semi-transparent volcanic ash clouds, where the cloud top height retrievals become sensitive to other reflective surrounding surfaces (water/ice clouds) the detection of accurate volcanic ash cloud heights is limited.

- Fresh volcanic layers are typically rich in water vapor (volcanic clouds also contain high concentrations of water). Due to this fact, the classification in the CALIPSO vertical feature mask sometimes fails to pick up the volcanic ash or sulfate aerosol because of competing clouds. TROPOMI may underestimate actual ash heights in case of semi-transparent volcanic ash clouds, especially in the presence of high concentrations of water vapor and for very high-altitude volcanic ash clouds (Hedelt et al., 2019, 2021).

Additionally, we provide an example below, showing the CALIOP VFM (left) and cloud phase (right) images corresponding to the Sinabung eruption. Types of clouds are flagged in the released VFM as cirrus - ice clouds. The “detected feature” is marked with a dashed red circle.



Minor comments:

Line 26: What does "3 and 4±3km" mean?

It refers to 3±3 and 4±3 km; this was altered accordingly.

Line 27: Correlation coefficients?

Of course, you are right. Updated in the text.

Lines 50-52: The sentence is too long and difficult to follow. There are several other places where shorter sentences may help the readers.

Thank you, we have split this sentence into two.

Line 59: What do you mean by "direct validation"?

Direct is a validation/verification that compares two products directly, on a one-to-one basis. The word is added to show the difference to the validation/verification performed via the CAMS assimilation experiments where the S5P SO2 LH is assimilated, and the resulting LH is compared to the IASI ULB/LATMOS LH. I.e. in an indirect way.

Line 79: "has been kicked-off" should be "was kicked off".

Agreed, changed.

Line 98: " By thus" should be "By"?

Agreed, changed.

Line 179: These are the conditions under which a retrieval would be considered valid for comparison? Please clarify.

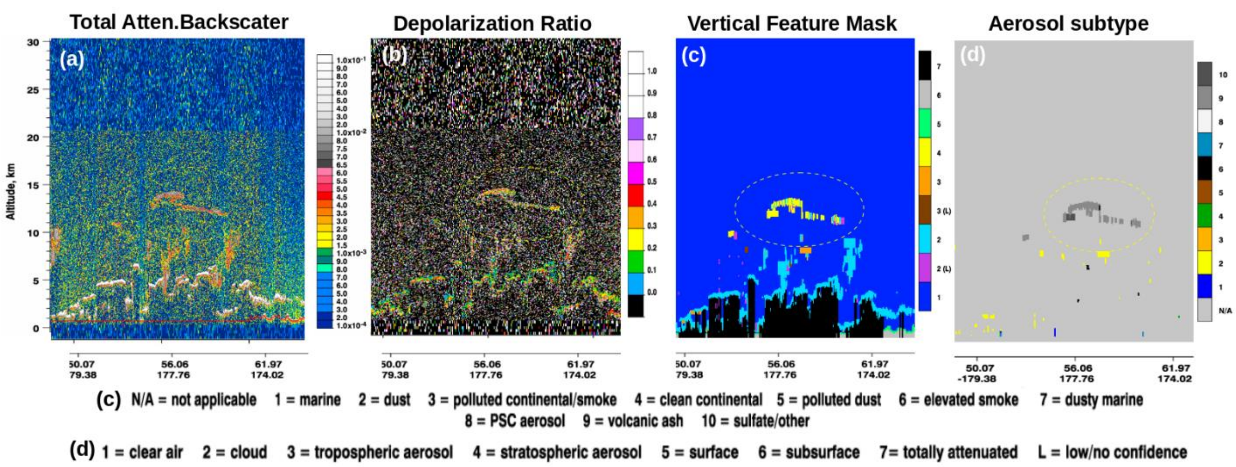
Thank you, it was indeed a confusing point. We have rephrased the entire two final paragraphs.

Line 189: what are the "three modules"? Please specify or remove the statement, otherwise it could be confusing.

The note on the “three modules” was rephrased to “three major algorithm steps” referring to the CALIPSO algorithm processing steps, in the updated manuscript. Briefly:

The level 2 processing involves three major steps. First, cloud and aerosol layers are identified by a set of algorithms, applied to the 532-nm attenuated backscatter profiles. After this, a set of scene classification algorithms (SCA) classifies these layers by type. Using data from the CALIOP channels, layers are identified as clouds or aerosols and the aerosol type and then the cloud ice–water phase are determined. At the end, profiles of particle backscatter and extinction coefficients are retrieved by the extinction retrieval algorithm performing retrievals within the layer boundaries identified before. Finally, classification procedures then proceed, layer-by-layer.

In the figure below we illustrate an example (not added in the revised manuscript) for 25 June 2019 for Raikoke. An uplifted ash layer can be seen using the CALIOP level 2 products. The (a) panel shows vertically resolved 532 nm attenuated backscatter [km-1sr-1], (b) panel show corresponding depolarization ratio, (c) the vertical feature mask and finally the (d) display the Aerosol subtype of the scene. Depolarization measurements indicate that the aerosol in the layer was predominantly spherical and therefore comprised mostly of sulfate.



Line 284: "a kilometre high" refers to the height or thickness of the ash plume?

The thickness, indeed. We have rephrased.

Line 370 (also lines 380-382): "satisfactory" - means the difference is within the expected uncertainty range of TROPOMI retrievals?

We consider that, especially in the comparisons with the ash plume observations by CALIPSO, the main source of uncertainty is the fact that the two plumes do not always coincide, as the separate – typically- early on after eruption. Hence, an overall difference of 2km can be considered spectacular keeping in mind of course that individual days' comparisons vary. In the new, colour-coded per eruptive day, scatter plot in Figure 8, one can note that for most of the days the comparisons have a small spread. There are however some, for e.g. the 28th and 29th, with numerous collocative pixels, that show a higher spread. No clear time evolution of the differences could unfortunately be identified.

Line 388: "a heights" should be "heights" or "altitudes".

Agreed, changed.

Lines 403-405: Perhaps briefly explain the physical processes that cause the bias in TROPOMI retrievals? I assume retrievals are possible with thin clouds above or below the volcanic plume, but only possible with bright clouds below the plume?

Our results highlight that there is added value in study scenes characterized by complexity. We expect that in case of sufficiently ash amount, the cloud height data products provide accurate volcanic ash heights. In case of semi-transparent volcanic ash clouds, where the cloud height retrievals become sensitive to other reflective surfaces below transparent volcanic ash clouds, detection of accurate volcanic ash heights is limited. This is very crucial in most of cases included in our analysis.

The low bias of the TROPOMI-CALIPSO comparisons may be expected as CALIPSO observes the top plume height and TROPOMI observes an average plume height when multiple layers are present. The CALIOP instrument furthermore possibly underestimateσ the aerosol layer thickness due to strong attenuation of the lidar signal at the top of the aerosol layer whereas TROPOMI SO2LH product does not suffer from such attenuation.

Of course, all these are on top of the major fact that the ash and SO_2 plumes do not always coincide to begin with.

Figure 11: Perhaps indicate in the figure caption that middle panels are for CAMS without assimilating TROPOMI retrievals and lower panels are with assimilation.

Agreed, changed.