Answer to anonymous referee #3:

In this paper, authors presented results from validation of GOME-2 Absorbing Aerosol Height (AAH) product against aerosol layer height from CALIOP Vertical Feature Mask(VFM) product for a selected suite of volcano cases. The objective of this paper is clear, which intends to present the performance evaluation of GOME-2 AAH product. However, there are several issues and challenges related to validation, which authors themselves also identified as well.
First of all, AAH from GOME-2 represents centroid height of absorbing aerosol layer, while CALIOP identifies the height of each detected aerosol layer, so, it is not clear that how GOME-2 performs for a single layer vs. multi-layers of aerosols presented in the atmosphere. In another words, it is strongly suggested to authors to clarify how the maximum and minimum height is derived from CALIOP for two situations and how GOME-2 performs.

Added to the manuscript to clarify this:

"As long as the altitude difference between the different layers provided by the VFM product was less than 200 m, the different layers were considered as one big layer in this study. For these layers, the minimum (minC) and maximum (maxC) height was then determined. This was done to reduce the amount of CALIOP data to compare with GOME-2 overpasses."

Secondly, the authors' intention is to validate the performance of GOME-2 AAH for volcanic ash, which is the reason that the validation cases are specifically selected for volcanic eruptions. However, as noted by authors as well, CALIOP have troubles to give the correct type for volcanic ash, therefore, the analyses of GOME-2 performance for different aerosol types identified by CALIOP, seems to me, do not have any merit, and the type actually pre-defined due to case selection.

As stated, our study focuses on the quantitative validation of the GOME-2 AAH product for cases of volcanic eruptions. It is therefore natural that respective cases were selected. In section 2.2, the CALIOP Vertical Feature Mask (VFM) is described, with references describing the capabilities and uncertainties of the derivation of aerosol layers and types. Within these limitations, the aerosol type derivation by CALIOP can be regarded as correct. Although our study focuses on GOME-2 AAH and volcanic ash, we think that it is worthwhile to present also the performance for the AAH not only when looking at all heights for all aerosol types but also for the different types. It is already mentioned that these analyses have to be taken with caution (cf. lines 341-342). We will add, however, some text to state this more clearly:
in line 297:
"Note, that the results for different aerosol types have to be analysed with caution, as the data were retrieved specifically for periods of volcanic eruptions".
And also in the supplement, in line 8:
"Note, that the results for different aerosol types have to be analysed with caution, as the data were retrieved for periods of volcanic eruptions."

Thirdly, the authors claimed that the performance of GOME-2 AAH does not have dependence on the distance and time in the matchups. This seems to be not convincing, since the analyses were performed when both factors are tangled together.

Please see our answer to this point further below (changes to figures 2 and 3).

Some comments and questions are given as follows.
Major comments:
1. Figure 1 and subsequent figures shows one AAH values corresponds to multiples layer height values from CALIOP, it is evident that this is caused by the criteria used for matchups. So, instead of plot all points, why the authors cannot plot mean value and standard deviation from CALIOP for each GOME-2 AAH values? And also those large outliers are from matchups with a large distance difference or a large time difference? It is worth to investigate...

Figure 1 has been adjusted according to the reviewer's comments and now shows for each GOME-2 AAH value the mean and standard deviation of the corresponding minimum CALIOP layer heights.

The outliers are discussed in the text (lines 281-286):
"In Fig. 1 there are points for which the average CALIOP minimum layer height is higher than 12 km and the corresponding GOME-2 AAH is much lower (< 3 km for GOME-2A and < 9 km for GOME-2B). For GOME-2A, most of the corresponding individual CALIOP pixels (85 %) were classified as volcanic ash, sulfate or elevated smoke layers and are classified by GOME-2 as pixels with high reliability. For GOME-2B however, only 28% of the corresponding individual CALIOP pixels were classified as stratospheric aerosol species but 95% of the GOME-2B pixels have a medium or low reliability level."

Changes to the manuscript:

New figure 1:

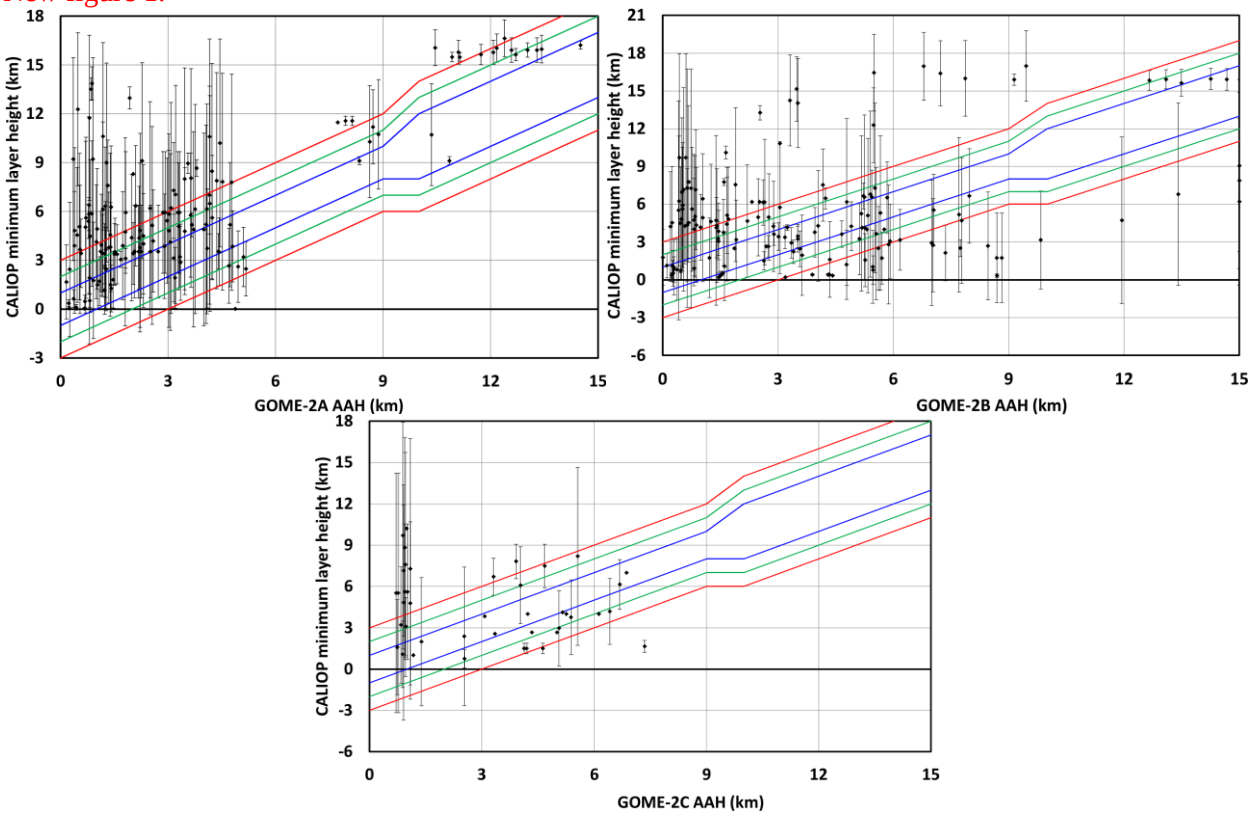

Fig.1: Requirement plots for GOME-2A (upper left), GOME-2B (upper right) and GOME-2C (lower middle). The red, green and blue line represent the threshold, target and optimal requirement lines. For each GOME AAH, the corresponding mean minimum CALIOP layer height and the standard deviation are shown.

2. Analyses of Figure 2 and 3 seems have two factors tangled together. To clearly demonstrate the claim that the degree of agreement between GOME-2 and CALIOP does not have dependence on the difference in both space and time, the authors should bin one variable when analyzing the variability of the other variable.

We created a new figure that combines the information from both figures 2 and 3, using colors to represent the different time difference bins.
As the results do not change, the adaptions in the text (paragraph lines 287-295) are limited to referring to Fig. 2 only and to delete Fig. 3 references.

Changes to the manuscript:

New figure 2: (+ removed figure 3)

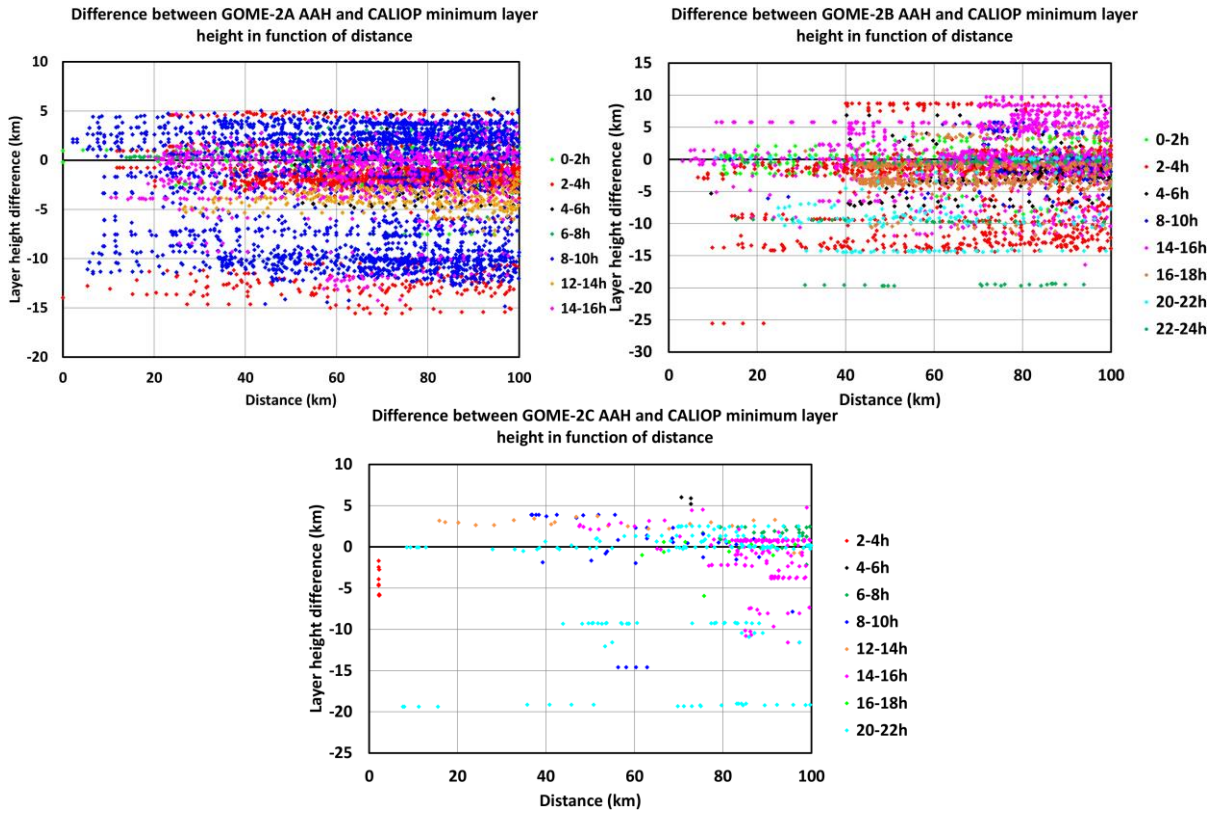

Fig. 2: Difference between GOME-2 AAH and the minimum CALIOP layer height in function of the distance between the GOME-2 and CALIOP pixel. The different colors represent different classes of time differences between the GOME-2 and CALIOP overpasses. The upper left plot shows the results for GOME-2A, the upper right plot shows the results for GOME-2B and the lower middle plot shows the results for GOME-2C.

3. It is understandable that matchup of GOME-2 with CALIOP is challenging, but the matchup criteria between GOME-2 and CALIOP might be too loose, which may lead to these large scatters in the results. Authors are encouraged to explore a relative tight criteria, it is ok to have less matchups, but do not want to include the matchups which really smear the results.

We studied the impact of using stricter criteria both for the maximum distance and for the time difference between GOME-2 and CALIOP overpasses separately.
First, we lowered the maximum distance from 100 km to 50 km. This reduced the dataset significantly, as only about 20%, 23% and 12% of the original data remained for GOME-2A, -2B and -2C respectively. When looking at the amount of data corresponding to the accuracy requirements, we found that by applying this tighter distance threshold, the amount of data in the threshold, target and optimal categories was reduced. In the scatterplot, there still remained a cloud of points with low AAH and high CALIOP minimum layer heights. This cloud represented 24%, 21% and 20% of the remaining data set for GOME-2A, -2B and -2C respectively.
Then, we kept the original distance limit of 100 km but we applied a threshold on the time between overpasses and only selected overpasses within 6h of each other. This also reduced the dataset: about 50%, 40% and 3% of the original data remained for GOME-2A, -2B and -2C respectively. For GOME-2A and -2B, the amount of data agreeing with the threshold and target requirements increased. For GOME-2C, too little data remained and the amount of data in the threshold, target and optimal requirements decreased to 23%, 8% and 0% respectively. Also for the reduced time difference, there still remained a cloud of points with low AAH and high CALIOP minimum layer heights. This cloud represented 21% and 20% of the remaining data set for GOME-2A and -2B respectively.

We decided to keep the original analysis with no limit on the time between GOME-2 and CALIOP overpasses and with a limit of 100 km for the maximum distance between overpasses. Our data set at the moment is too reduced by putting more strict thresholds in place and the improvement in the statistics is too small to make the change worthwhile. In the future when more data can be included in our analysis, we will revisit the option to impose stricter thresholds.