We thank both reviewers for their kind words about our work and for their comments/suggestions. Please find their original text in italic with our answers below.

Anonymous referee #1

General comments

The identification of parameters that influence the data quality of atmospheric trace gas retrievals is a central question in many validation analyses. But it poses intrinsic challenges due to the multi-dimensional and non-linear nature of the problem. The authors of "Using self organising maps to explore ozone profile validation results – SCIAMACHY limb compared to ground-based lidar observations" propose a machine learning method, based on self-organising maps (SOMs), to uncover patterns in the quality of the retrieved profile data. They explain their method step-by-step, along with illustrations for the particular case of SCIAMACHY limb ozone profile data which is compared to lidar observations. The paper goes beyond a mere description of the method, as it continues with an interpretation of the results that lead to several conclusions on the variables that influence SCIAMACHY limb ozone profile data quality.

The proposed SOM-based method is novel and generally applicable to any comparison study of two vertical trace gas profile data sets, and therefore relevant for a wide audience in atmospheric sciences. In addition, the authors succeeded in providing a concise yet pedagogical description of their method, which should be accessible for readers not acquainted with machine learning methods. The scientific relevance and presentation quality are hence very high. However, the scientific quality should be improved before accepting this paper. The results shown for the illustrated case are not sufficient to motivate the author's conclusions on the relationship of various parameters with SCIAMACHY ozone data quality. Even stronger, I am concerned that the interpretation is at least partially wrong, as I explain in the specific comments below. When higher-dimensional and/or non- linear problems are tackled with sophisticated and powerful methods the user risks over-interpreting the results if no basic cross-checks are made, e.g. using traditional methods. Since this paper may open the door for many readers to start using SOMs as well, it is imperative that the authors provide an exemplary illustration not only of the method, but also of the subsequent interpretation of the results. The latter is not the case, according to me. I therefore recommend a major revision, in which the authors either provide further evidence in support of their claims, or refrain from the discussion of the interpretation of the SCIAMACHY results.

Specific comments

R: *My main objection is the interpretation of the results in Section 3.4 (p4384 l25 – p4385 l12). Since these are repeated quite prominently in the abstract (p4375, L10-16) and in the conclusions (p 4388, L25 – p 4389, L13) these form an important part of the paper. In find that the authors interpret the altitude-dependent correlation of the explanatory variables (Figure 7) in a careless way. They state that with increasing altitude the quality of SCIAMACHY limb ozone (between 18 km to 36 km) depends mostly on longitude (lon), then solar zenith angle (SZA), then latitude (lat) and solar azimuth angle (SAA). The patterns in Figure 7 for these four parameters are similar, and in addition, it is quite clear from Figure 6 that they all correlate very well with the lidar station. The latter is also pointed out by the authors to some extent, in Section 3.3 (p4384 L2-6). According to me, there is therefore a more natural explanation for the patterns in Figure 7, namely that the codebook vectors correlate mostly with lidar station (which can unfortunately not be computed). In this case, the patterns in Figure 7 are a result of the correlation of the four EVs with the station's location. To sort this out, it would be instructive to show*

- Correlation plot of codebook vector versus (lat,lon,SZA,SAA) at various altitude levels, where each marker is coloured according to the dominant lidar site.

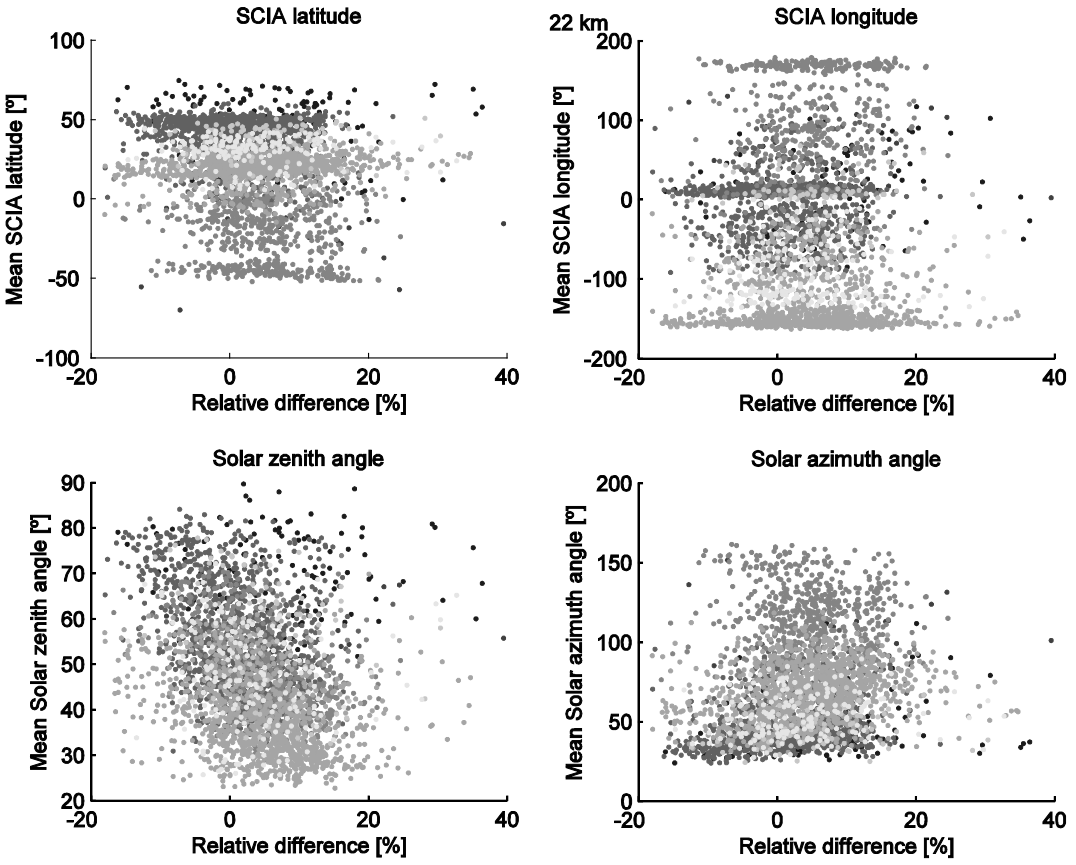

Figure R1. Cross plot of de-normalised codebook vectors [%] with representative EV values [º] for SCIAMACHY latitude (upper left), SCIAMACHY longitude (upper right), solar zenith angle (lower left) and solar azimuth angle (lower right) at an altitude of 22 km.

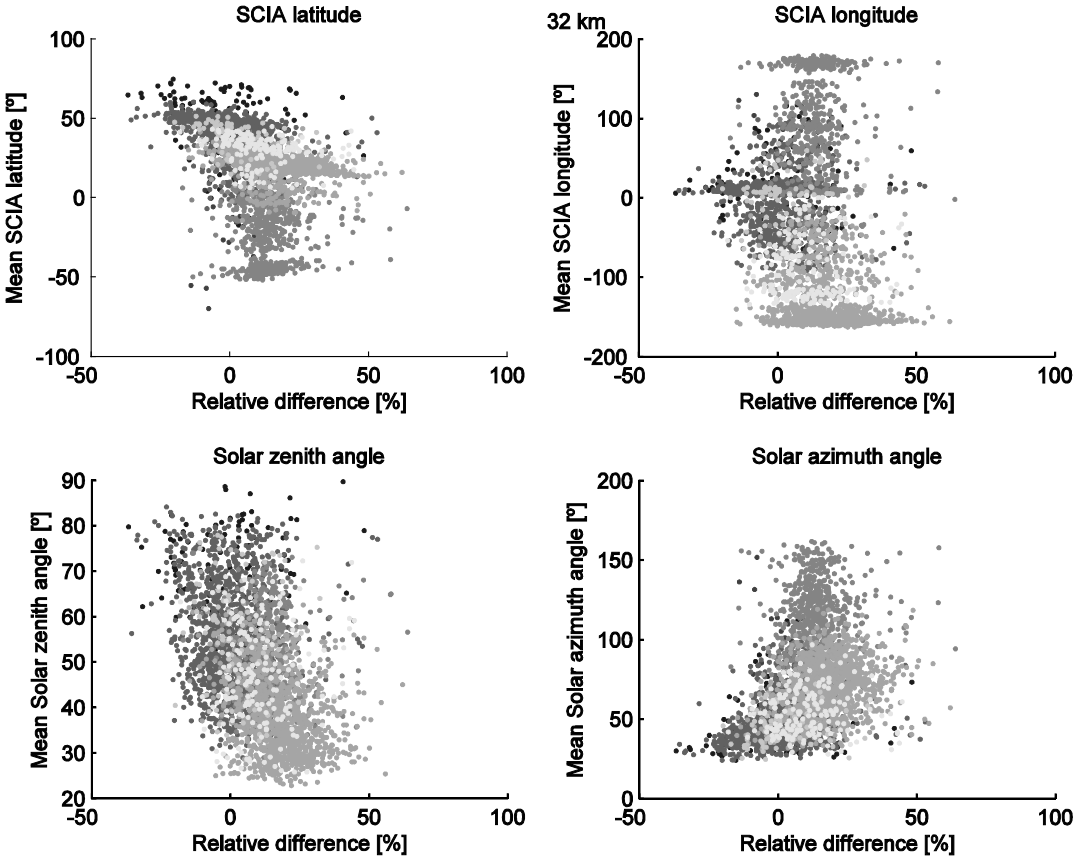

Figure R2. As Fig. R1 but for an altitude of 32 km.

- At the beginning of the analysis, a plot of the altitude-dependence of the (median/mean +- spread) normalized differences at each of the seven lidar sites.

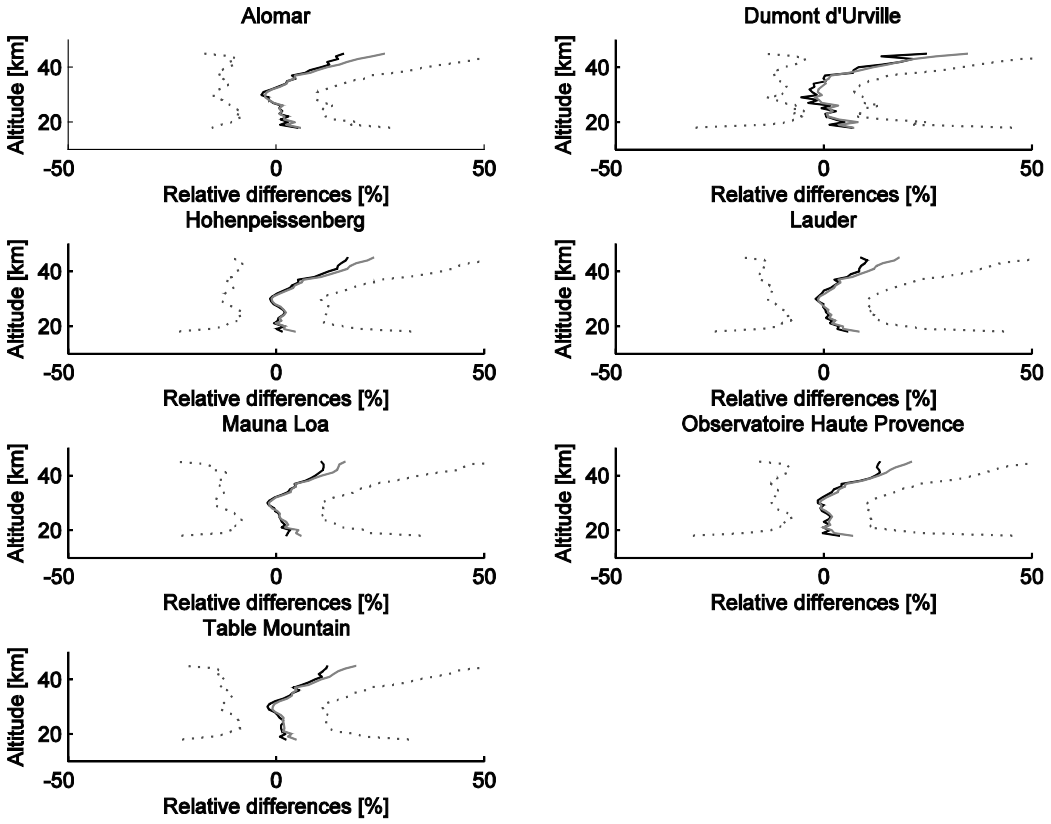

Figure R3. Relative differences between SCIAMACHY and lidar for the seven collocation sites as a function of altitude. The black line corresponds to the median, the dark grey line to the mean and the dotted lines to the mean plus/minus one standard deviation.

- A correlation plot of each pair of lat, lon, SZA and SAA would help, with markers coloured by lidar site.

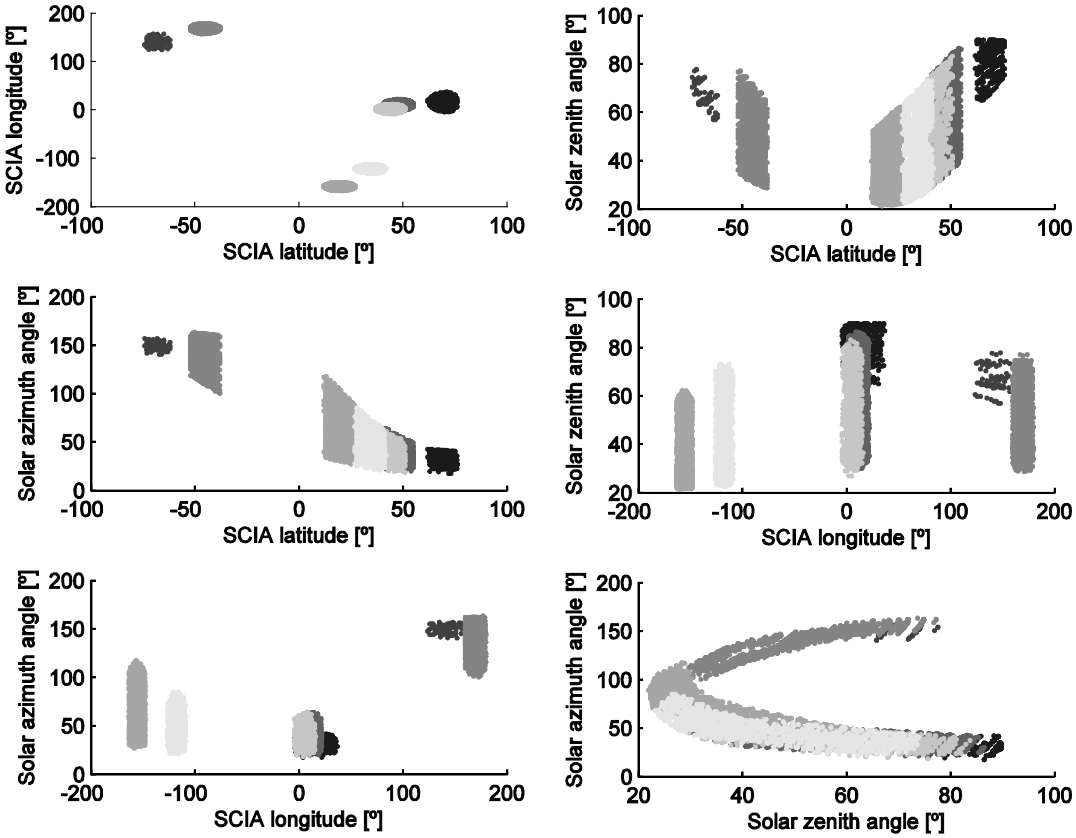

Figure R4. Cross plots of all combinations of the selected EVs (SCIAMACHY longitude, latitude, solar zenith angle, solar azimuth angle) for the input data.

R: *If these plots do not provide support for the author's interpretation, the corresponding parts of the paper need revision. In a similar way, Figure 10 may be overinterpreted (Section 3.4 and the conclusions), and a simpler fundamental explanation is possible.*
A: We have used the best available dataset, which unfortunately does not optimally sample all ranges of EVs as the amount of available sites is limited and data quality filtering further reduced the number of collocations. Correlation between the EVs, such as solar zenith angle and latitude is inherent due to the SCIAMACHY measurement pattern. However, as can be seen in Figs. R1, R2 and R4, the relations are not 1 to 1 and most of the variation in EVs is still being captured by multiple sites. Specifically, in Fig. R4, the top right plot shows the latitude versus the solar zenith angle. Most of the variation in solar zenith angles is covered by at five stations, with only the most extreme cases being covered by just one single site. We thus conclude that although the range in EVs may be limited and correlated to some extent through the selection of the ground stations and of SCIAMACHY pixels around these sites, we deem the covered variation sufficient to use it to study the dependencies between EVs and the observed ozone profile differences. We will amplify the discussion in the manuscript.

R: *A second important concern is the representativeness of the SOM training sample. The definition of the training sample is a very important step in any machine-learning based-method. This should be stressed out more in the paper, and I would welcome some more motivations behind the choices made by the authors. I suspect that the comparison sample is not homogeneously distributed over the lidar stations. Figure 6 e.g. shows that three sites dominate most neurons (Hohenpeissenberg, then Mauna Loa, and finally Lauder). Is it possible that the bulk of the data in the training sample corresponds to these stations? Could you e.g. add the number of pairs per site in Table 1? I do not see how one can arrive at representative conclusions with a non-representative input sample.*
Coming back to the remarks in the previous paragraph: it is possible that this is at the heart of the patterns seen in Figure 7. How can you be sure that SCIAMACHY limb ozone data quality truly depends on lat, lon, SZA and SAA? The training sample could be made more representative by restricting the same number of pairs per lidar site. And even then, it is not a trivial task for the specific example given

the limited collocation statistics of satellite-ground comparisons. The method therefore seems to have more potential for satellite inter-comparisons where, in some cases, much higher collocation statistics can be gathered and more representative training samples can be obtained.
A: In machine learning, results can only be as good as the (training) data is. Thus the representativeness of the data is, indeed, of great importance to get good results and the data that goes into the SOM will define what information can be extracted from it. Here we have taken all the data that was available at the time and which fulfilled the quality/selection criteria as outlined in section 3.1. We would like to stress that with SOMs, in contrast to other neural network approaches the 'training' dataset is not a subset of the data, but all the available data was used to train the SOM (optimize the values of the codebook vectors in each SOM unit – neuron).
The samples are not homogenously distributed among the test sites, with especially the number of valid collocations at Dumont d'Urville being very limited. However, this is not a problem for our SOM-based method as multiple samples can map to the same SOM unit so we do not need to sub-sample our dataset. For instance, if a very large number of observations for a particular site behave in a similar fashion, these will map to the same SOM neuron or to a small set of contiguous neurons. In traditional analyses such an uneven distribution of samples bias the results as these observations would heavily influence the outcome of the correlation if the training data is not sub-sampled.
The three sites mentioned indeed have the largest amount of samples. They are prominent in Fig. 6 because we labelled each neuron with the name of the site that had most observations mapping to it. A more detailed analysis of the sites that map to each neuron shows that 61% of the neurons has more than one site mapping to it. Figure R5 shows the second most present site at each neuron. The more homogeneous distribution shows that the three mentioned sites are not dominating the analysis.

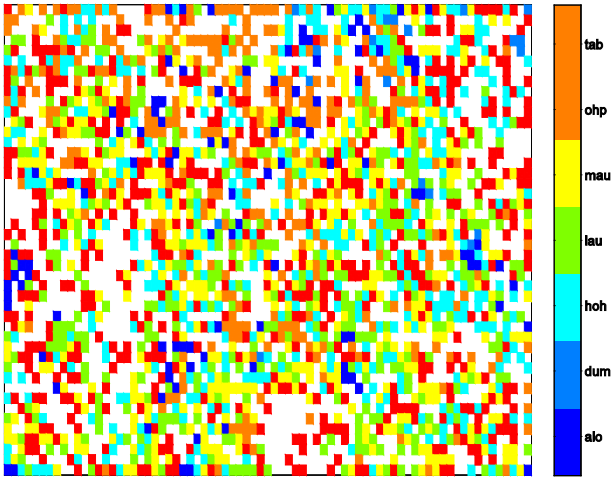

Figure R5: Map of the second largest contributing site to each SOM neuron

Technical corrections
T: *P4375 L17: exploring → exploration*
A: The text will be updated.
T: *P4375 L21: coupling → merging*
A: The text will be updated.
T: *P4376 L5: inter-cluster → inter-class (?)*
A: The text will be updated.
T: *P4376 L21: remove "used"*
A: in our view it is not necessary to remove 'used'.
T: *P4377 L1: increased → increases*
A: The text will be updated.
T: *P4377 L5: proposed → considered*
A: The text will be updated.
T: *P4377 L15: Sect. 2 → Section 2*
A: Author instructions state that Section/Figure etc. need to be abbreviated when not at the beginning of the sentence.

T *P4377 L21: remove "proposed"*
A: The text will be updated.
T : *P4378 L1: SCIAMACHY "limb"*
A: The text will be updated.
T: *P4378 L3: launched "into space"*
A: The text will be updated.
T: *P4378 L17: remove "which is"*
A: The text will be updated.
T: *P4379 L11: increasing → degrading*
A: The text will be updated.
T: *P4380 L2: are → is*
A: The text will be updated.
T: *P4380 L4-8: how exactly is the mapping done? How is the "distance" defined between a difference profile and a codebook vector?*
A: For the mapping, we first calculated mapping indices for each data sample. For this, we found the SOM unit (neuron) that minimizes the Euclidean distance between each the data sample (normalized difference profile) and the codebook vector "stored" in each SOM unit.
T: *P4381 L10: not at all → not all*
A: The text will be updated.
T: *P4381 L14-15: put "with the matching metadata" between brackets*
A: The text will be updated.
T: *P4386 L15-16: remove phrase "The codebook vectors … by the indices". It is a repeat of previous paragraph.*
A: The text will be updated.
T: *P4387 Last paragraph: double check the naming of cluster 1, 2 and 3. Do they really correspond to what we see in Figure 10? It appears to me that it is the other way around.*
A: Indeed, this is incorrect. The clusters were ordered the other way around in a previous version. Thanks for catching this mistake.
T: *P4388 L15: After that the role → After that, the role*
A: The text will be updated.
T: *Table 1 Please add number of collocations per site.*
A: The table will be updated.
T: *Figure 1 L2: trained used a → trained on a*
A: The text will be updated.
T: *Figure 1 L3: the SOM consists of "a" grid formed*
A: The text will be updated.
T: *Figure 5 Add unit [%] in caption. Please clarify in the main text or in the caption whether the orientation is identical as for Figure 6 and Figure 9. I assume they are, but the aspect ratio suggests the opposite.*
A: As mentioned in the caption, the maps are stretched vertically for a better visualization. The orientation is thus the same. The unit will be added to the caption (is currently implied with 'relative').
T: *Figure 7 Reduce the number of labels on the Y-axis.*
A: The number of y-tick labels will be reduced.
T: *Figure 9 Clarify that orientation is as in Fig 5 & 6.*
A: Orientation is as in Fig. 5 and 6.
T: *Figure 10 Is the naming of the clusters the same as in the main text?*
A: Thanks for pointing out this mistake in the document. In a previous version of the figure the clusters were ordered differently.

Anonymous referee #2

General remarks:

The paper describes a methodology to apply the mathematical technics of self-organizing maps (SOMs) to explore ozone profile validation results of SCIAMACHY limb compared to ground-based lidar observations of 7 different NDACC-stations. The aim is to study the influence of a set of different observational characteristics such as ge- ographical location, solar zenith angle etc. In itself an interesting study and could be appropriate for publication in AMTD. Although the major content of the paper is well structurized and written, it still suffering that essential information and critical analysis of the data is missing. The use of SOMs in the comparison of different data sets is not that new and since more than 10-20 years well known and also well established in atmospheric sciences, e.g http://www.cis.hut.fi/projects/somtoolbox/. This is also the SOMs-toolbox the authors have used for their study.

A: The toolbox is indeed not new, but we did not find any publication where SOMs are used in an analysis such as proposed here, especially the aspect of mapping the EVs onto a trained SOM.

R: *In the present form the paper describes only the methodology of applying SOM's in mathematical terms, however, the compilation of mathematics into physical relevant terms is rather poor. Particularly in chapter 3 the physical base is almost completely missing.*
The paper should contain a more precise and thorough analysis based on interpretation in physical terms of the obtained results.
In addition, the final results should also be (quantitatively) discussed in comparison with results obtained by other investigators using traditional validation methods.

A: No journal publications exist for the SCIAMACHY version 5.02 limb ozone profiles nor for the previous data version. We have identified various dependencies between the EVs and observed ozone profile differences at different altitudes. This information can be used by the algorithm developers to try to attribute these observations to the corresponding areas in the retrieval algorithm. For instance, the relation with latitude may originate from the apriori profile, or be related to straylight, et cetera. The approach used here with the selected EVs cannot identify the exact sources of the deviations in the retrieved profiles, but does hint where to look for improvements of the retrieval model.

R: *Further, it seems that the number of vertical profiles of each Lidar station applied in the analysis are not equally distributed in time and space and so introducing artefacts which are finally dominating the results of the SOMs-analysis such that consequently the conclusions made are neither adequate. This would mean that the analysis and the interpretation of the results has to be re-done. Therefore, I rate the paper as being only acceptable for publication after major revisions as also written in my specific comments below.*

A: See answers to reviewer 1. The nature of the method actually avoids introducing artifacts by the non-homogeneous distribution of input data in space and time. Of course, the results that are shown are influenced by the availability of the data to train the SOM. We deem the variability present in the data sufficient for inferential analysis. In fact, we have used all available data in the network fulfilling the quality criteria. Therefore it is important to sustain and possibly introduce additional validation sites that further extend the range of the key explanatory variables. Such an analysis could be used to confirm the results presented here.

Specific Comments:

R: *Below I have listed my major points of critics. Most essential is that the compilation of SOM-mathematics into physical terms is very poor and mostly missing and certainly not discussed adequately*

Section 2.3: Physical base is missing

A: We are not sure what the reviewer means by 'physical base', so we translate the SOM terms to domain terminology. For instance, a codebook is a vector of representative values for the set of normalized relative difference profiles that map to that neuron. Similarly, a component plane is a matrix with all the representative values –codebook elements- at a given altitude.

R: *Section 3.1: Missing: number of lidar profiles per station finally used. In how far ho-mogeneity of Lidar and Sciamachy vertical profiles is a pre-requisite? Should number of*

difference profiles from each lidar station be the same and identical distributed in space and time?
A: The unevenness of the distribution of the observations per site was also raised by the first reviewer. As explained in our answers above the number of profiles per station is not required to be the same. In fact, this method actually minimises the impact of having an uneven distribution of number of collocations per site. For instance, if a very large number of observations for a particular site behave the same, these will map to a small set of contiguous neurons whereas in traditional analyses they would heavily influence the outcome of the correlation. Homogeneity of the vertical profiles is not a pre-requisite as if e.g. at one location, the profiles are very distinct from the rest, those profiles would be grouped together adjacently on a part of the SOM with the rest mapping away from that cluster.

R: *Page 4381, Line 14: What do you mean with matching meta data. Please explain in physical terms.*
A: With matching metadata is meant: the EV values corresponding to the collocated measurements of the difference profile. For example, the difference in equivalent latitude between the lidar and SCIAMACHY observation.

R: *Section 3.2: Physical base is missing. P4382,L02: Why 46x75 and why hexagonal neurons?*
A: In all SOM-based studies, the number of neurons and the size of the SOM map is arbitrarily decided by the user. A smaller SOM map is typically used when SOMs are used for clustering (grouping various profiles into one or more contiguous neurons). In our case, we do not want to cluster the data so we used a large number of neurons. In fact, we selected this SOM size by setting that on average four profiles map to one neuron. A hexagonal neuron lattice is a preferred SOM cell format because it has six direct neighbours and neither the horizontal nor the vertical direction is favoured (Kohonen, 2001). Again, it is an arbitrary choice. Using a square neuron lattice, a slightly different number of neurons/map size, or a different map form (e.g. toroid) will not change the conclusions substantially.

R: *P4382,L03: What are "input vectors" in physical terms.*
A: An input vector is a vector of the normalized differences for one collocation pair that goes into the SOM.

R: *Figure 5: What are the horizontal and vertical axis representing?*
A: The axes represent the two dimensions of the SOM.

R: *P4382,L25: What are "codebook vectors" in physical terms.*
A: The codebook vectors are the representative normalized relative differences.

R: *Section 3.3: P4383,L12: The explanatory variables (EV) should be here discussed in more detail. What kind of meta-data information for each EV has been used in the mapping the EV-planes. This is essential in order to interprete the outcoming results.*
A: The metadata corresponds to the EVs associated with the normalized differences that map to each neuron. In other words, the information comes from the characteristics of the set of observations that map to a neuron. The mapping is done minimizing the Euclidean distance between the relative ozone differences profile and all of the codebook vectors, therewith finding the most similar representative differences profile.

R: *Figure 6: a.) What are the horizontal and vertical axes representing?*
A: The axes are the same as in Figure 5 (which was stretched vertically) and the image thus covers the SOM.

R: *Figure 6: b.) The pattern of the "Location" map is dominated by HOH, MAU and LAU which is also seen in "Latitude" and "Longitude" map. Less clearly but still visible a similar pattern is seen in the "Solar zenith angle" and "Solar azimuth angle". All other maps show no significant pattern. Knowing that OHP is located very close to HOH I would expect both station represented more equally*.
A: Figure R6 shows where all Hohenpeissenberg and OHP input vectors have mapped onto the

SOM (presence indicated with a black pixel). The images for these two neighbouring sites clearly show similar patterns.

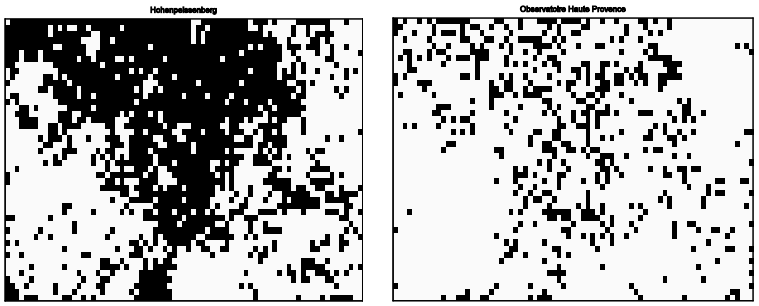

Figure R6. Presence mask for Hohenpeissenberg (left) and Observatoire Haute Provence (right). A black pixels indicates that at least one input vector from the mentioned site mapped to that neuron.

R: *It seems that the number of vertical profiles of each lidar station applied in the analysis are not equally distributed in time and space and so introducing artefacts which are dominating the results of the SOMs-analysis. In general the question is raised: Does it make sense to have the seven lidar stations together with latitude and longitude as three independent EVs? Latitude and longitude are not really independent parameters but pre-dominated and directly linked to the geographical locations (incl. matching criteria of 800 km) of the 7 Lidar sites?*
A: Please see the answers above and to reviewer 1. It would be nice to increase the number of validation sites, but that can only be achieved by having a less strict data selection or by reducing the altitude coverage.

R*: Section 3.5: P4385,L21: Explain k-means in physical terms? P4386: What is the physical meaning of the 3 clusters obtained and what they represent?*
A: K-means is a popular machine-learning algorithm here used to group together the differences of ozone profiles based on their similarity. Each cluster then represents a group of profiles characterized by some features or subtle differences (e.g. cluster 1 has a fewer pronounced differences at the highest altitudes than the other 2 groups).

T: *Chapter 4: change title "Summary and Conclusions".*
A: We will update the text.