Interactive comment on "Application of High-Dimensional Fuzzy K-means Cluster Analysis to CALIOP/CALIPSO Version 4.1 Cloud-Aerosol Discrimination" *by* Shan Zeng et al.

Anonymous Referee #2

Received and published: 28 September 2018

The manuscript describes a methodology to discriminate between aerosol and cloud layers from CALIOP/CALIPSO lidar Level 2 data based on the high dimensional Fuzzy K-Means Cluster Analysis. The argument for sure is a good fit for the journal but some parts are not clear, probably suffering from hasty writing and need improvements before final publication. Moreover, other tests should be performed to improve scientific significance and clarity. I am however confident that the authors will brilliantly address all the issues I raised.

Thanks for reviewing the paper and giving valuable feedback. It is very hard to validate the operational algorithm at global scale, because we know of no existing global in-situ data set that could be used for the task. A comparison between different classification schemes used by active and passive sensor has been done in previous work (Stubenrauch et al., 2013). However, as active sensors profile the full vertical extent of the atmosphere, it remains quite difficult to compare classification results with passive sensors that, at best, only measure the properties of a single layer. (More often, properties of multiple layers are convolved into a single set of measurements, and thus tasks such as separately classifying cirrus clouds and boundary layer aerosols within the same pixel are extremely challenging retrievals for passive sensors.) Furthermore, comparisons between different algorithms have not yet been performed. Similar to the comparison between passive and active sensors, it's hard to determine how accurate the algorithms are (see our previous comments about the use of synthetic data), but by combining data from multiple sensors we can estimate upper and lower boundaries for cloud and aerosol distributions over the globe, and these values give a distribution range to guide modelers. Similarly, the comparison between supervised and unsupervised algorithms can also give upper and lower boundaries for precision to guide modelers, instrument developers, and data processors. To address the referee's concerns, we add detailed statements in the introduction to clarify these points. As we say in the conclusions of the original draft, the purpose of this study is "to validate the performance of the cloud-aerosol discrimination (CAD) algorithm used in the standard processing", and we are not suggesting FKM as a replacement for COCA. To this end we also added more introduction about the importance of discriminating between clouds and aerosols, and described the benefits for the study for different user communities.

Major Comments:

The FKM clustering methodology is well described and totally makes sense. But, as stated in the introduction, the FKM method is used to validate the result of V4 CAD algorithm and to better understand the classification, identifying the crucial parameters. It looks like that all the produced efforts have a very low return on investment. The V4 CAD is not validated vs. a reference dataset, i.e. using a synthetic lidar data where all the aerosol and cloud properties are well known and controlled, but with respect to another methodology that have comparable uncertainties.

While the use of synthetic data as an evaluation tool is generally an excellent and highly effective strategy, in the case of discriminating clouds from aerosols it's also especially hard to implement in a useful way. For ~90% of the cases, cloud and aerosol properties are very well separated and reliable classifications can be made using a single wavelength elastic backscatter lidar (e.g., CATS; see https://cats.gsfc.nasa.gov/media/docs/CATS_QS_L2O_Layer_3.00.pdf). In these cases, unambiguous synthetic data can be used to weed out those algorithms that are obviously deficient. But the remaining cases fall into the cloud-aerosol overlap region (see Liu et al., 2009), and for these layers even the most sophisticated human observers cannot always agree on the correct partitioning (e.g., see Koren et al., 2007; Tackett and Di Girolamo, 2009; Varnai and Marshak, 2011; Balmes and Fu, 2018). These especially difficult cases include separating thin cirrus from lofted Asian dusts, separating evaporating water cloud filaments from the surrounding aerosols in the marine boundary layer, and separating fresh volcanic ash from cirrus. Given the measurements available on the CALIPSO platform, the classification of these targets is always subject to some uncertainty. So, yes, we certainly could create "synthetic lidar data where all the aerosol and cloud properties are well known and controlled" and compare the classifications obtained from the CALIPSO operational CAD algorithm (COCA) and the FKM algorithm. And by using this synthetic data to compare algorithm outputs versus "truth" we could perhaps choose an algorithm that best confirms our own prejudices; but whether that algorithm was actually delivering the correct classifications in the really hard cases would still be an open question.

Moreover, It is completely missing an analysis on who is really using those data, i.e. climatologists, modelers..., and why it is critical to discriminate (defining a level of precision) between aerosols and clouds (and their subtypes). For example, how much is it the actual precision of the current operational V4 CAD algorithm in classifying the aerosol and cloud layers ? The final users are ok with this accuracy? Which benefits will be obtained reducing the misclassification? How the FKM will be used or implemented to reduce the V4 CAD misclassification?

While this would certainly be interesting information, this kind of detailed analysis lies well beyond the scope of this paper. (Simply counting up the number of different CALIPSO data user communities that make use of the CAD scores we provide would likely lead to some fascinating (and perhaps surprising!) insights.) In this paper, our goal is limited to providing a performance assessment of the current CALIPSO operational CAD algorithm.

In the manuscript is only marginally discussed why January 2008 measurement are a representative data sample. How the results are impacted changing the analyzed dataset?

As one-month data is enough for the purpose of the study, we just randomly choose one month. For different dataset, the class number and the fuzzy exponent may be different, but classification results on cloud and aerosol should not be too different in theory. In reality, for different season, different features occur which may slightly impact the sample of classes and thus the results. The paper just focuses on the first step of comparison and didn't go further. We mentioned this in the data preparation and added a summary of future work at the end of conclusion.

The number of classes is predefined (2 or 3) after analyzing Figure 3. However, in operational contexts, some data subsets might belong only to two classes. FKM still will fill with observation the class that should be empty. Is there a reason why the authors used the FKM cluster analysis instead of some self-selecting class methods, i.e. MeanShift clustering (Cheng, Yizong. "Mean shift, mode seeking, and clustering." IEEE transactions on pattern analysis and machine intelligence 17.8 (1995): 790-799) or classification algorithms as AdaBoost (Hu, Weiming, Wei Hu, and Steve Maybank.

"AdaBoost-based algorithm for network intrusion detection." IEEE Transactions on Sys- tems, Man, and Cybernetics, Part B (Cybernetics) 38.2 (2008): 577-583) ?

I think both Meanshift and Adaboost are very good algorithms to do clustering too. There are so many clustering methods (more than 100 maybe), supervised or unsupervised, connectivity-based, centroid based, density based and distribution clustering, we only try the Fuzzy K-means out, which is one of unsupervised centroid method that produces a membership which (between 0 and 1) is represent the probability of belonging to one class and is comparable to the official CAD scores (between -100 to 100, probability belong to one and the other). And also the shape of multi-dimensional observations of cloud and aerosol are suitable for centroid based algorithms. Last, the FKM unsupervised approach is quite different from the highly supervised method used to train the operational algorithm, is what we need for the comparison and the objective of the study.

Density based algorithms such as Meanshift expect some kind of density drop to detect cluster borders. Mean-shift is usually slower than k-Means. Besides that, the applicability of the mean-shift algorithm to multidimensional data is hindered by the unsmooth behavior of the kernel density estimate, which results in over-fragmentation of cluster tails (Achert et al. 2006). Clouds have two centers (ice and water) and aerosols may also have several sub-centers (e.g., dust and biomass burning), so a density based algorithm may not suitable for this classification in my opinion. Also, according to Kaur and Chawla (2015), FCM has higher accuracy compared to the Meanshift. AdaBoost is a machine learning method, and more complicated to understand. While using it may resolve the problem for FKM weighting problems in some future study, at the moment we want an easier understand method that is distinctly different from the COCA method investigate different algorithm inputs on the classifications. In the future we will consider to doing some machine learning classifications, but may not choose AdaBoost.

The random initialization of the centroids is a well-known problem as the initial centroid selection not only influences the efficiency of the algorithm, but also the number of relative iterations (and consequently the needed time machine). Some optimal centroid selection techniques can be found in Nazeer, K.A. Sebastian, M.P Clustering biological data using enhanced k-means algorithm". In: Electronic Engineering and Computing Technology, Springer, 2010, pp. 433–442 (chapter 37)

The flowchart is wrong in previous version of manuscript. We have a loop to choose the best initiation and outcome results in FKM algorithm. With the loop to choose the best initiation, the larger the number of loops, the better the resulting clusters will be, but this is not time efficient. In application to real data, we have not yet found that using a larger number of loops will consistently improve the classification accuracies for the CALIPSO level 2 observations.

Many thanks to the reviewer for introducing us to an efficient way to save the relative iteration number and time.

Specific comments:

Line 27 Pag. 1 Please add also "geometrical properties"

We added it.

Line 15 Pag. 5 How the random initialization influence the final result? I don't recall any section where this issue is discussed. Are the results consistent with the random initialization?

We misrepresented our algorithm, and so we modified our flowchart in Figure 1. As we do a loop to choose the best random initialization, outcome results do not change due to initiation as long as the iteration number and the loop number for selecting initiation are big enough.

Line 16 Pag. 5 the authors mean Equations 2, 3 and 4?

We corrected them.

Figure 1: Third step it should be Eq. 6 and 7

We corrected them.

Line 2 Pag. 7: I am not sure that latitude is not useful to discriminate, as clouds at 16 km at polar latitudes may rise a flag, as cirrus clouds below 9km in the equatorial and tropical regions

The region (i.e. latitude) and season information are of course useful auxiliary information because they can indicate the sources of particles and the dynamics of the atmosphere. The others are directly measured optical information of the particles due to their scattering nature. In the future, we can train and apply the FKM method at local scales, which could be a way to improve the current classifications.

Figure 3: labels are difficult to read. The picture in the middle shows "NCE" that is not previously defined.

We selected the bold font to the labels so as to see the label easier and changed the "NCE" to "MPE".

Line 14 Pag 11: please rephrase "water clouds. For these water clouds".

We rephrased it.

Figure 4: it is very hard to see the zone of interest (smoke and cloud). Maybe reduce the vertical scale from 0 to 20 km?

We modified it to 20km.

Line 15 Pag 17 please read "We saw" instead of "We see"

We corrected it.

Paragraphs 3.4 a, 3.4 b and 3.4 c. How the authors assume that the layer are pure dust, smoke and ash respectively? Is there any other ancillary measurement that shows without any doubt the aerosol layer composition?

This comment highlights one of the major difficulties in validating a global data set acquired by a first-of-its-kind active sensor: coincident measurements of interesting phenomena are extremely difficult to come by! For these events, we tracked these plumes by eye according to the event's location, time period and our experience in evaluating spatial distributions and layer optical features (depolarization, color ratio and backscatter). This is very accurate though.

Section 4. Figure 13 is not very intuitive and it is difficult to get meaningful information from it . It might be interesting to replace it (or add) the Screen Plot and the loading factors as barplot as showed in https://doi.org/10.1175/JTECH-D-15-0085.1.

The figure includes a lot of information compared to the barplot, but we did not explain it well. We have now added more explanation about the figures and added the color bar.

Line 4 Pag. 34: Even if the FKM Cluster Analysis closely replicate the CAD V4 operational algorithm, it is not validate it (see main comment section)

We changed the "validation" to "comparison". We explained more in the paper that the comparison between algorithms can set up boundaries for the uncertainness due to different algorithms

Line 18 Pag. 35. FKM it is a time consuming algorithm because setting up random centroids can slow down the convergence process and in some cases can produce as result sub-optimal centroids virtual centroids (i.e. not corresponding to any observational measurement). See Main Comments section.

Yes, we added more details to the related domain to clarify the reason for FKM "time consuming". We modify the algorithm description in the paper and in Figure 1.

References

Achtert, E.; Böhm, C.; Kröger, P. (2006). "DeLi-Clu: Boosting Robustness, Completeness, Usability, and Efficiency of Hierarchical Clustering by a Closest Pair Ranking". LNCS: Advances in Knowledge Discovery and Data Mining. Lecture Notes in Computer Science. 3918: 119–128. doi:10.1007/11731139_16. ISBN 978-3-540-33206-0.

Balmes, K. A., and Fu, Q.: An Investigation of Optically Very Thin Ice Clouds from Ground-Based ARM Raman Lidars, Atmosphere, 9, 445; doi:10.3390/atmos9110445, 2018.

Brakhasi, F., Matkan, A., Hajeb, M., and Khoshelham, K.: Atmospheric scene classification using CALIPSO spaceborne lidar measurements in the Middle East and North Africa (MENA), and India, International Journal of Applied Earth Observation and Geoinformation, 73, 721-735, https://doi.org/10.1016/j.jag.2018.07.017, 2018.

Kaur, S. and Chawla S. (2015): Evaluation of Performance of Fuzzy C Means and Mean Shift based Segmentation for Multi-Spectral Images, International Journal of Computer Applications 120, 25-28.

Koren, I., Remer, L. A., Kaufman, Y. J., Rudich, Y., and Martins, J. V.: On the twilight zone between clouds and aerosols, Geophys. Res. Lett., 34, L08805, doi:10.1029/2007GL029253, 2007.

Liu, Z., Vaughan, M., Winker, D., Kittaka, C., Getzewich, B., Kuehn, R., Omar, A., Powell, K., Trepte, C., and Hostetler, C.: The CALIPSO lidar cloud and aerosol discrimination: Version 2 algorithm and initial assessment of performance, J. Atmos. Oceanic Technol., 26, 1198-1213, https://doi.org/10.1175/2009JTECHA1229.1, 2009.

Liu, Z., Kar, J., Zeng, S., Tackett, J., Vaughan, M., Avery, M., Pelon, J., Getzewich, B., Lee, K.-P., Magill, B., Omar, A., Lucker, P., Trepte, C., and Winker, D.: Discriminating Between Clouds and Aerosols in

the CALIOP Version 4.1 Data Products, Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2018-190, in review, 2018.

Tackett, J. L., and Di Girolamo, L.: Enhanced aerosol backscatter adjacent to tropical trade wind clouds revealed by satellite-based lidar, Geophys. Res. Lett., 36, L14804, https://doi.org/10.1029/2009GL039264, 2009.

Varnai, T. and Marshak, A.: Global CALIPSO Observations of Aerosol Changes Near Clouds, IEEE Geosci. Remote Sens. Lett., 8, 19-23, https://doi.org/10.1109/LGRS.2010.2049982, 2011.