Atmos. Meas. Tech. Discuss.,
doi:10.5194/amt-2018-168-RC2, 2018

Atmospheric Measurement Techniques
Discussions

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.

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. 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?

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 ?

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 Systems, Man, and Cybernetics, Part B (Cybernetics) 38.2 (2008): 577-583) ?

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)

Specific comments:

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

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?

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

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

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

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

Line 14 Pag 8: please rephrase "water clouds. For these water clouds".

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?

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

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?

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 Scree Plot and the loading

factors as barplot as showed in https://doi.org/10.1175/JTECH-D-15-0085.1.

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)

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.

Interactive comment on Atmos. Meas. Tech. Discuss., doi:10.5194/amt-2018-168, 2018.