Geoscientific Model Development Discussions

Geosci. Model Dev. Discuss., referee comment RC2
https://doi.org/10.5194/gmd-2021-191-RC2, 2021

Comment on gmd-2021-191

Anonymous Referee #2

Referee comment on "An aerosol classification scheme for global simulations using the K-means machine learning method" by Jingmin Li et al., Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2021-191-RC2, 2021

Dear authors

Thank you for submitting this work for review. Firstly, I would agree that it is interesting to evaluate methods for grouping model outputs into distinct clusters, if only for the purposes of an additional diagnostic. As we know, aerosol formulations within large scale models are not ideal; largely as a result of the computational challanges and chemical complexity retained within earth system modules. We should not shy away from presenting results that are negative in the sense of clearly demonstrating limitations of proposed methdologies, however simple or complex they are. All this aside, I do feel this paper requires more work and clarity before it could be published in GMD as per the following discussion.

Whilst K-means is a fast technique, it dopes rely on a number of assumptions on the data in focus. This includes the presence of clusters with equal size and density. Limitations also include sensitivity to outliers and poor accuracy scaling with increasing dimensions. There are studies that modify k-means to increase its performance, includingthe use of neural network techniques that act as a non-linear dimension reduction approach to generate K-means 'friendly' spaces. Have you used PCA to assess any changes in cluster properties? Presently there is not enough information in the manuscript on the distribution of each metric used in the clustering. Plotting the distribution of values for each metric would help guide the reader to better consider the appropriate choice of pre-processing. I understand that choice of clustering technique can be at the whim of the investigator, but there should be a level of data exploration that helps form the narrative around the relevance of the results.

With this in mind, I would like to see more discussion on the benefits and limitations of general unsupervised techniques earlier on in the manuscript. You do note these at the end of the paper, but I feel these should be discussed earlier.in the summary and outlook, you state that 'K-means has been proven to be a powerful classification tool' - is this with regards to this study or generally? Please clarify statements like this. Given the known

limitations, it is difficult to agree with this.

If K-means is used primiarly due to its computational performance, please state this with a discussion on the data challenges you have. It seems you do not have a significant amount of gridded data [~18K points?] to cause a problem with regards to computational cost. However, you mention 'the huge amount' of data from global simulations toward the latter stages of the manuscript. How many points did you begin with?

Following on from this, I would suggest you provide comparison with another method for clustering before reaching a set of conclusions as to the viability of K means. This could be hierarchical agglomerative clustering [HCA], with appropriate pre-processing as per the approach used with K means. If your dataset is indeed ~18K points this will not take long to compute. If this is a concern you may find significant improvements from the fastcluster package which can be called using the same syntax as those within Scipy: http://danifold.net/fastcluster.html. Just to re-iterate here: If you wish to demonstrate the performance of K-means then the justification and limitations, according to the volume and properties of your data, must be clear. Comparison with HCA, for example, may give this study a useful balance.

Please also add the balance of limitations in the abstract. The statement 'A markedly wide application potential of the classification procedure is identified and further aerosol studies are proposed which could benefit from this classification.' requires that additional body of work.

Please also provide the file used to perform the clustering. I have reviewed the zenodo instances for both model output and cluster script, but there seems to be a significant disconnect between the global model output store and a file used in the cluster procedure. Please provide at least so information on how one can extract the relevant files to ensure reproducability.