Geoscientific Model Development Discussions

Geosci. Model Dev. Discuss., referee comment RC1
https://doi.org/10.5194/gmd-2021-191-RC1, 2021

Comment on gmd-2021-191

Anonymous Referee #1

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-RC1, 2021

Review of Li et al.

Li et al. present research exploring the application of the K-means clustering method to simulated aerosol data. The authors claim that this method allows for the identification of aerosol regimes and demonstrate the relationship between these regimes and known aerosol sources and property distributions. While the application of clustering methods like K-means show great potential in exploring atmospheric composition, I cannot recommend this paper for publication at this time due to major methodological flaws and a lack of novel scientifically valid results. Major issues are summarized below.

Major Issues

Data Standardization

Gaussian standardization as described on line 210 is inappropriate for much of the data used here. Aerosol number and mass concentrations typically vary logarithmically (e.g. Figure 1a), and standardization should reflect this variability. Improper standardization can lead to spurious clustering and limits the interpretability of the results.

K Value Selection

The authors well describe two metrics for selecting the value of K, or number of clusters in the manuscript. However, the quantitative metrics are ignored in favor of "expert judgement" (e.g. line 266) when selecting the number of clusters in section 3.1 and 3.2. In both sections, the SSE plots (Figures 2a and 4a) do now show a distinct elbow and the maximal silhouette coefficient is at 2 clusters. Both of these figures indicate that in the standardized dataset used here there are no strong natural clusters and the applicability of the K-means algorithm should be revisited entirely.

Scientific Results

After applying the k-means clustering (with the major limitations outlined above) the authors do not find novel or (in some cases) scientifically valid results. Despite the claim that "specific aerosol characteristics for the predominate regimes are not known a priori" (line 91), a large body of research exists quantifying the regimes and characteristics of aerosol in various regions. The results of this work are at best consistent with that prior knowledge (e.g. the importance of emissions for controlling aerosol regimes in the lower atmosphere). In other cases, the results strongly disagree with prior knowledge in ways that are not adequately addressed in the text.

For example, in Figure 2 the clustering analysis shows that nearly all of western Europe is in a "background continental" regime, despite several major anthropogenic aerosol sources (i.e. Paris, London, Benelux). In the same clustering analysis, largely remote areas of Asia and the Middle East are classified in the same aerosol regime as Los Angeles, California, despite the very large differences in aerosol characteristics that are known over these areas. Additionally, the major dust source of the Gobi Desert is not present in any of the dust clusters. The large differences in the results from this manuscript and the existing literature on aerosol regimes potentially indicate larger issues in the results of the clustering algorithm application and are not addressed in sufficient detail in the manuscript.