Atmospheric Chemistry and Physics Discussions

Atmos. Chem. Phys. Discuss., referee comment RC1
https://doi.org/10.5194/acp-2021-126-RC1, 2021

Review of acp-2021-126

Anonymous Referee #1

Referee comment on "Identifying source regions of air masses sampled at the tropical high-altitude site of Chacaltaya using WRF-FLEXPART and cluster analysis" by Diego Aliaga et al., Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2021-126-RC1, 2021

The current paper describes potential origins of air masses arriving at the Chacaltaya (CHC) atmospheric research station using high-resolution numerical weather predictions from the weather research and forcasting model (WRF), back-trajectories from the WRF version of the Lagrangian particle dispersion model FLEXPART, and the K-means clustering algorithm. The power of the method is illustrated using a straightforward example. It is a thorough and well-written manuscript which can be useful for analysis of data recorded at the location of CHC in the future. As such, I am happy to recommend publication when the comment below has been adressed.

The most difficult part in analysing results from any clustering algorithm often is the selection of the number of clusters. The work could be nuanced here as in this type of numerical algorithms, the robustness and performance should be the main driver in chosing this parameter. I believe this choice should be adressed more carefully.

The selection of 18 clusters by the authors is founded by two reasons. The first is that it represents a local maximum in the parameter scan of the silhouette average score. The second uses prior assumptions of the authors based on the interest in 2 vertical levels, 2 horizontal scales, and 4 wind directions which led to the expectation of a solution near 2 x 2 x 4 = 16 clusters. Although this reasoning is very intuitive, I wonder if it is not too reductive in nature.
For example, it implies that all atmospheric observatories should be able to identify about 16 clusters when performing a similar analysis, independant of their location.
Furthermore, with equally valid reasoning one can caluclate a preference to other numbers of clusters. For example, the identification of 6 pathways (equivalent to directions in the near-field), would suggest a solution at 2 x 2 x 6 = 24 clusters. This would prompt the further investigation of the 23 cluster solution which shows a local maximum in the silhouette average score. Alternatively, one can assume that the vertical levels are coupled to the horizontal scales as wind speed in the free troposphere is generally larger than those in the boundary layer. This coupling does not allow multiplication to figure out the number of combinations possible. The result would not be 2 x 2 = 4 spatial ranges but e.g. 3: short range, medium range, and long range clusters (as found by the authors

when analysing the 18 clusters). Using the 4 directions from the assumption, you would expect a solution at 3 x 4 = 12 clusters, using the 6 major pathways instead of directions one would anticipate a solution at 3 x 6 = 18 clusters.

The 18 cluster solution, as selected by the authors, is good because it adequately described the data and has a straightforward interpretation. If the authors think this is insufficient reasoning to select the number of clusters, they can apply alternative clustering algorithms to illustrate the robustness of their choice. In my opinion, no further reasoning should be brought forward when using only 1 clustering technique as it will inevitably be subject to speculation.