Anonymous Referee #2

Received and published: 3 June 2018

Note regarding document formatting: black text shows original referee comment, blue text shows author response, and red text shows quoted manuscript text. Changes to manuscript text are shown as *italicized and underlined*. Bracketed comment numbers (e.g. [R1.1]) were added for clarity. All line numbers refer to discussion/review manuscript.

[R2.0] This manuscript discusses application of Hierarchical Agglomerative Clustering (HAC) to analysis of data collected using the Wideband Integrated Bioaerosol Sensor (WIBS4A). While real-time detection of bioaerosols has been quite well controlled, the analysis and classification is still challenging and vital problem. Therefore, investigation and improvements in this area are very important and crucial for understanding the abilities and limitations of LIF aerosol detectors. The manuscript is well written and in detail reveals important problems of fluorescence data analysis of bioaerosols. I recommend presented manuscript to publication, however some corrections and further explanations to the following remarks will be appreciated:

[A2.0] Author response: We thank the referee for her/his positive summary of the manuscript and recommendation to publish after comments are addressed.

[R2.1] 1. The techniques of single particle detection using LIF devices, like WIBS, reached relatively high reliability and perfection. The device collects data in real time, on the other hand the presented results are offline. The data analysis takes a long time. Finally, the standard methods like particle collection on tape is still competitive with LIF. My question is: Did the authors try or are going to apply real-time aerosol data analysis?

[A2.1] I think the statement that "LIF devices … reached relatively high reliability and perfection" is already an very optimistic statement, but I agree that when operated and analyzed properly the data can often be useful. The referee's suggestion about real-time data analysis is an interesting idea that has been discussed. We are working on this type of analysis from a different angle and with respect to a different class of instruments, but we have not had the ability to investigate real-time analysis strategies with respect to WIBS data. This would be a worthwhile project, but is outside the scope of what we were aiming to accomplish in this study and would likely require dedicated project funding.

[R2.2] 2. L67 - principle or principal component analysis?

[A2.2] In this case the word "principal" is the correct one. I often get this word confused with "principle" and have to look up the definitions to make sure I'm correct.

[R2.3] 3. L116 – "The WIBS collects
3 channels of fluorescence intensity. . ..." – collect channels or collects fluorescence intensity in 3 channels?

[A2.3] This was indeed poor grammatical construction. The sentence has been changed to: "The WIBS collects *information about* ~~3 channels of~~ fluorescence intensity ~~information~~ *in three channels* …"

[R2.4] 4. L170 – ". . .both saturating and non-fluorescent particles were retained. . ." – Did authors collect the particles?

[A2.4] We did not physically collect the particles. The wording here was unfortunately confusing. In this case we have "retained" the data in the analysis process by not removing particles based on certain attributes. To clarify, the word "retained" was changed to "analyzed" as shown here: "*… both saturating and non-fluorescent particles were analyzed* ~~retained~~ …"

[R2.5] 5. L370 – ". . .gains. . ." or grains?

[A2.5] This is a typo; "gains" was corrected to "*grains*".

[R2.6] 6. L494 - ..fluorescence and non-fluorescent particles.. - The phenomenon should not be compared with the property.

[A2.6] This typo was changed for the discussion version of the manuscript to be "fluorescent and non-fluorescent particles."

[R2.7] 7. L 424 and further – I think that term "synthetic mixtures" for recorded numerical data is confusing and should be corrected. Firstly, it sounds like a chemical synthesis process. Secondly, the final result of clustering should be the same and independent whether the particle data are sorted or not. Otherwise, the order (sequence) of detected particles would change final result. I think that actual meaning of used data is well described in L298-300 ("...subset taken from the pool of particles..".

[A2.7] The term "synthetic mixtures" is indeed confusing terminology, and this is a point raised also by Referee #3 (i.e. [R3.1], [R3.3], and [R3.6]). Referee #3 suggested the term "computational simulations" or "simulated mixtures" among several possibilities, and we have changed the text in a variety of places through-out the manuscript to reflect this new terminology.

[R2.8] 8. L 426 – "analytically synthesized" – analysis has opposite meaning to synthesis should be corrected

[A2.8] Here the term was changed to "computationally simulated."

[R2.9] 9. L 428, 431, 434, 436, – ". . .mixture synthesized. . ." – see point 7.

[A2.9] The word "synthesized" was changed to "simulated" in each of these cases and all others within the manuscript.

[R2.10] 10. The authors compared clustering ability using selected small groups of substances. It would be interesting to see the clustering output for all 14 types together. Why it was not presented?

[A2.10] This additional experiment might be interesting, but it is unlikely to add anything to the general nature of the conclusions. The 14 types of particles assembled for these match-up experiments (i.e. Sections 4.1 – 4.3) were meant to be individually instructive, but not to represent the entirety of the types of particles one might see in a more complex, ambient environment. So collecting all 14 into one experiment would represent another experimental combination, but would in itself not be any more relevant than the individual simulations already discussed.