Anonymous Referee #1
Received and published: 1 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.

[R1.0] This paper builds on existing literature examining unsupervised learning techniques to improve the interpretation and classification of data obtained with WIBS UV-LIF spectrometers. As shown in previous publications, Hierarchical Agglomerative Clustering (HAC) can serve as a robust data analysis method for classification/interpretation of bioaerosol data but the accuracy of technique is highly sensitive to the choice of clustering linkage and data pre-treatment (e.g., Crawford et al., 2015); this is further explored in this paper which elucidates how data pre-treatment choices such as choice of fluorescent threshold and log normalising data may influence clustering accuracy using laboratory samples of known particle types (Savage et al., 2017) in various synthetic mixtures, and thus the authors present tentative recommendations of data pretreatment regimes depending on the analysis goals. Overall the paper is well written and the computational experiments well thought out. The findings here are useful and further validate the usefulness of Hierarchical Agglomerative Clustering for interpretation of WIBS data. The results also provide a useful framework for testing Hierarchical Agglomerative Clustering data pre-treatment regimes for other atmospheric science data problems and neatly demonstrate the potential pitfalls of not rigorously performing such tests. I recommend publication after the following comments have been addressed.

[A1.0] Author response: We thank the referee for her/his positive summary of the manuscript and recommendation to publish after comments are addressed.

Specific comments

[R1.1] L73-77: The authors have conflated some of the terminology relating to unsupervised and supervised leaning methods. I'm uncomfortable with the use of the term clustering when discussing supervised methods as clustering specifically relates to cluster analysis. I suggest replacing "clustering techniques" with "classification algorithms" and "(trains) the clustering algorithm" with "(trains) the classification algorithm".

[A1.1] The referee raises a good point. We changed terminology on page 2 according the referee suggestions, as listed below:

- L68: "*Classification algorithms, including several* clustering techniques in particular, have shown successful results …"
- L73: "~~Clustering techniques~~ *Classification algorithms* can be divided …"
- L76: "This type of method enhances (trains) the ~~clustering~~ *classification* algorithm in that the output ~~cluster classes~~ *groups* are predetermined …"

[R1.2] L120: Please state the bands and what they relate to.

[A1.2] Additional text was added, as shown below:
"The WIBS collects 3 channels of fluorescence intensity information (FL1, FL2, and FL3), particle size, and particle asymmetry for each interrogated particle. *The bands of excitation and fluorescence emission are: FL1 (λ_{ex} = 280 nm, λ_{em} = 310 – 400 nm), FL2 (λ_{ex} = 280 nm, λ_{em} = 420 – 650 nm), and FL3 (λ_{ex} = 370 nm, λ_{em} = 420 – 650 nm).* The excitation and emission wavelengths chosen for each of the 3 fluorescence channels were designed to maximize the

information gained about key biological fluorophores present in a broad range of bioparticles (Kaye et al., 2005; Pöhlker et al., 2012). *Early generations of UV-LIF bioaerosol spectrometers were often interpreted to be able to detect proteins via channels similar to FL1 and products of active cellular metabolism (i.e. riboflavin and NAD(P)H) via channels similar to FL3, but these approximations are gross simplifications that confound more detailed investigation of particle types.*"

[R1.3] L198: Can the authors please clarify why they have used log spaced bins. Do you mean that you have taken a log of the data and it is binned naturally by the discrete nature of the detector resolution (i.e., fine bins) or have you binned the data into specific (coarse) log bins? If it is the latter can you please state what the bins are and can you comment on how forcing the data to in bins may influence the clustering? My concern here is that too coarsely binning the data may create artificial hotspots due to reduced resolution and bias the clustering, reducing the capacity to differentiate between particles with similar properties. Can the authors comment on this and demonstrate the effect this may have by providing an example for comparison where the data is converted to log space and not binned. I also wonder if the bins should be normalised by the bin width to further complicate matters.

[A1.3] Aspects of this discussion are presented in L209-212. To summarize in different words, the data values from a given channel were either used as recorded (i.e. "value") or as logarithmically transformed (i.e. "log(value)"), depending on the Scenario. The values were not forced into specific bins, but rather input into the cluster algorithm using the exact value in either of these forms. The reason that logged values can provide different results by HAC is that the distance between points is different in linear space or log space, because the cluster process does not independently take into account whether a value is as recorded or as log(value). Because many real-world particle variables can present normal distributions only in log space (i.e. lognormal size distributions), we explored inputting values in both raw and log forms.

The following sentence was added to the manuscript at L211 for clarity:
"*By this process, data values were input into the algorithm as log(value), but without additional binning*."

[R1.4] L254: Can the authors comment on the environmental applicability of the chosen ratios. I would suspect that in an urban environment you may expect something closer to a ratio of 1:99 fungal to diesel particles with the converse being true in a forest environment. How does the clustering perform under such extreme mismatches?

[A1.4] We originally explored three different ratios of particle concentrations (80:20, 50:50, and 20:80) for each of the three match-ups discussed in Figure 3 in order to show that input ratio can be important to how the algorithm responds. This was certainly not intended to be exhaustive, and one could additionally explore more extreme ratios. So to limit the scope of the analysis here, we chose to present evidence only that the ratio matters, without trying in all cases to predict ratios that could be relevant to a wider range of ambient environments.

The question the referee brings up is interesting, however, and so we explored 1:99 ratios of each of the three particle type combinations presented in Figure 3, where the bioparticle is the minority concentration in each experiment. The results are shown below in a plot/table form identical to how they are presented in Figure 5. The Bacteria:Diesel and Fungi:Dust separations still performed quite well (6.6% and 13.5% misclassification, respectively), even with the extreme mismatch of input concentrations. The Fungi:Diesel separation was poor, however, in a 2-factor solution, because the Diesel particles split into both clusters, and the Fungi particles were likely

too low in concentration to influence the cluster properties. We added text including a summary of these new experiments to the manuscript at L304:

"To extend the investigation of particle input ratio, the three match-ups presented in Figure 3 were investigated using Scenario B with 1% bioparticles and 99% non-bioparticles in each respective case. In these experiments the Bacteria:Diesel and Fungi:Dust particles separated relatively well (6.6% and 13.5% misclassification, respectively). The Fungi:Diesel separation was poor, however, because the Diesel particles were nearly evenly split into both clusters, and the Fungi particles were too low in concentration to influence the cluster properties. More investigation is needed to explore how extreme disparities in particle ratio could negatively influence cluster quality in real-world settings."

Part A: Individual Clusters (Particle Number)

	Cluster	B3	F2	S4	D12
Fungi : Diesel	1	-	37	2588	-
	2	-	0	1111	-
Bacteria : Diesel	1	57	-	4	-
	2	0	-	5653	-
Fungi : Dust	1	-	45	-	7
	2	-	12	-	5650

Part B: Grouped Clusters (Particle Number)

	Cluster	Bio	Non-bio
Fungi : Diesel	1	37	2588
	2	0	1111
Bacteria : Diesel	1	57	4
	2	0	5653
Fungi : Dust	1	45	7
	2	12	5650

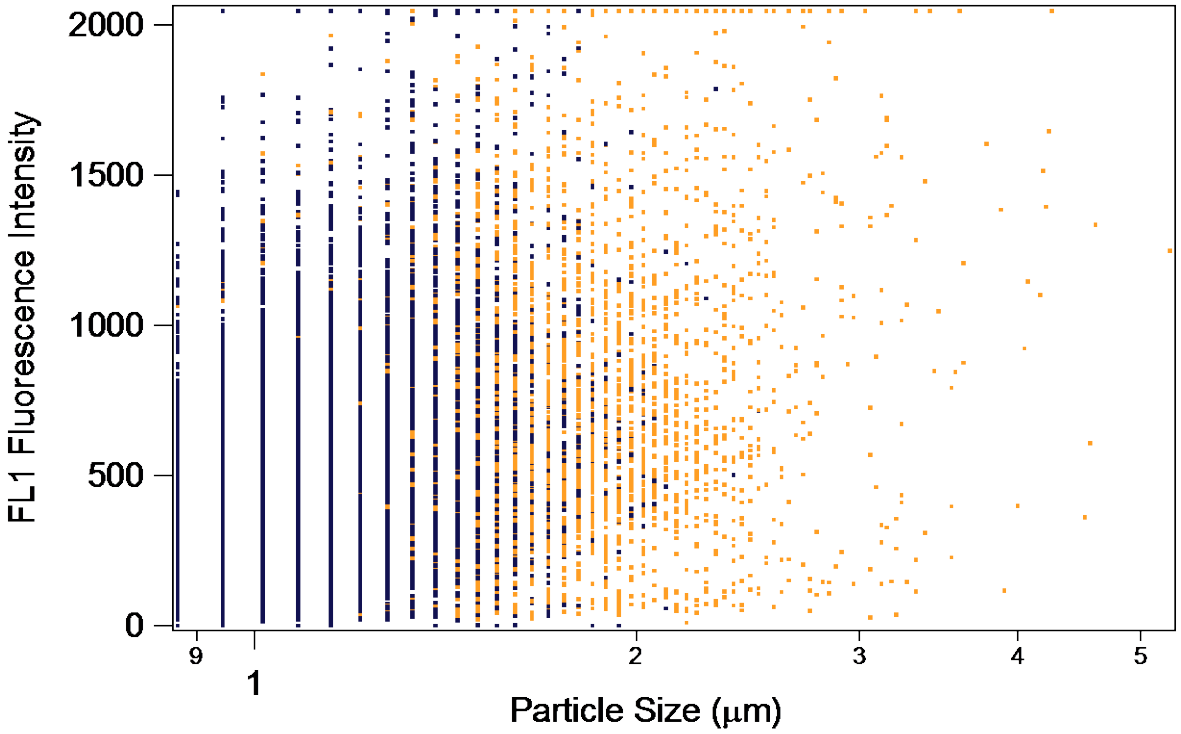
Part C: Summary (Cluster Quality)

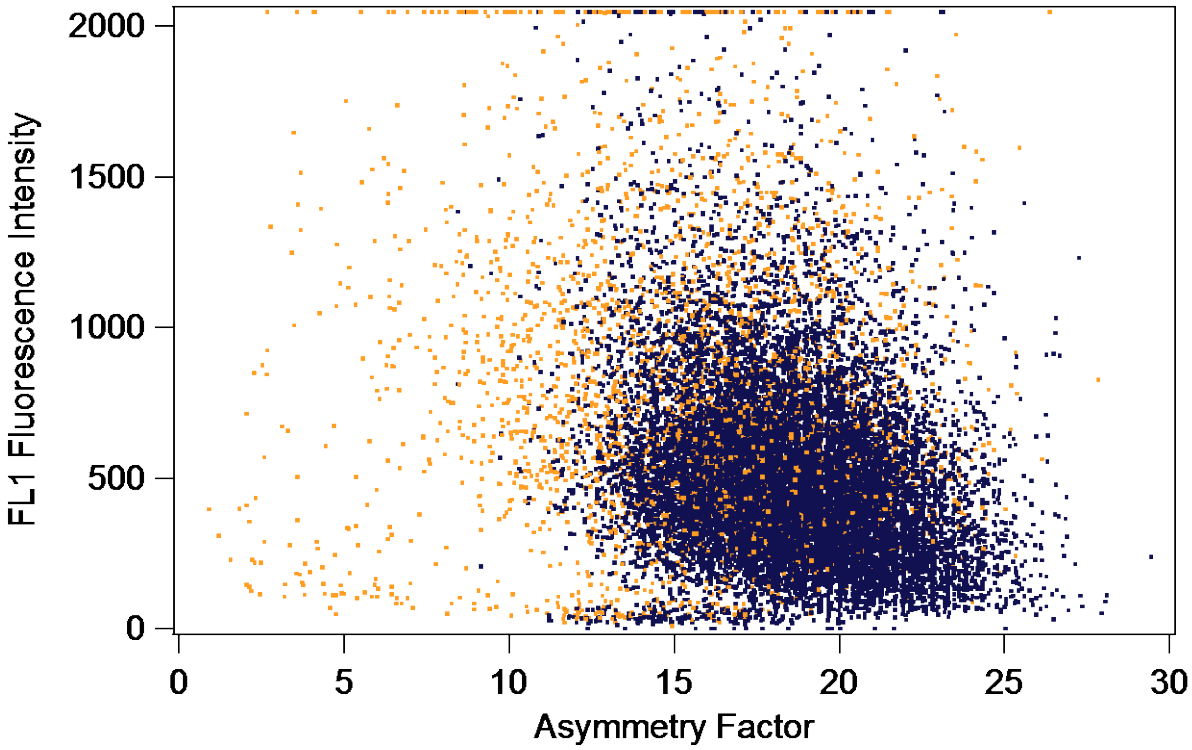
	Total P.	Miscl.	Cat.
Fungi : Diesel	2625	98.6%	Fungi
	1111	0.0%	Diesel
Bacteria : Diesel	61	6.6%	Bacteria
	5653	0.0%	Diesel
Fungi : Dust	52	13.5%	Fungi
	5662	0.2%	Dust

[R1.5] L238: Would it be possible to show examples of the cluster centroids for a case where there is significant misclassification? This may illuminate why the algorithm is failing to correctly attribute particles. It may also be useful to examine the fluorescence/AF characteristics of each cluster as a function of size here. A 2D histogram or color density plot could show distinct hot spots that haven't been separated correctly and could provide a basis for manual separation based on sensible thresholds.

[A1.5] To address the referee's suggestion, we included an additional set of plots here as suggested. The results below correspond to the match-up between Bacteria 1 and Bacteria 3 using Scenario B and the 3-sigma threshold, which corresponds to Experiment 22 from Table 2 (65% misclassification). The two colors of dots in the plots represent clusters 1 and 2. In this case it is still unclear how to utilize a single threshold to separate between the two particle types here.

In the process of analyzing results of this study we produced countless plots and tables, each of which showed slightly different angles of the same story. We chose to simplify the results in many cases to make the manuscript shorter and more manageably readable. We find that the table of fluorescence intensity and AF median values (Table 2 from original data published in Savage et al., 2017) often summarizes the differences in the particle types rather well and so were rarely able to separate using 2D histograms as the referee suggests. One example of these two additional plots is included here for reference, however.





[R1.6] L312-315: Can you describe the method for producing the soot as they seem rather large as compared to that in the study of Toprak and Schnaiter (2013) which were also coincidently found to be weakly fluorescent in FL1. Perhaps the soot used in this study is larger and more fluorescent than we may

expect of ambient/urban soot which may cause some of the difficulty in correctly attributing in in some cases?

[A1.6] The method for aerosolization of particle types discussed was presented in Section 3.2 of the associated Savage et al., 2017. Specifically, the aerosolization details related to soot are copied here:

> From Page 4284, Section 3.2.3 of Savage et al., 2017: "Dry powders were aerosolized by mechanically agitating material by one of several methods mentioned below and passing filtered air across a vial containing the powder. For each method, approximately 2.5–5.0 g of sample was placed in a 10 mL glass vial. For most samples (method P1), a stir bar was added, and the vial was placed on a magnetic stir plate. Two tubes were connected through the lid of the vial. The first tube connected a filter, allowing particle-free air to enter the vessel. The second tube connected the vial through approximately 33 cm of conductive tubing (0.25 in. inner diam.) to the WIBS for sample collection."

The referee is correct that the method of producing/aerosolizing particles, including soot, will bear heavily on the fluorescent properties observed. In particular, different aerosolization methods are likely to produce very different size distributions, which then will dictate the overall fluorescence properties. For this reason, we included the following statements in the Savage et al., 2017 paper:

> From Page 4292, Section 4.3: "It is important to note, however, that the method chosen for particle generation in the laboratory strongly impacts the size distribution of aerosolized particles. For example, higher concentrations of an aqueous suspension of particle material generally produce larger particles, and the mechanical force used to agitate powders or aerosolize bacteria can have strong influences on particle viability and physical agglomeration or fragmentation of the aerosol (Mainelis et al., 2005). So, while the absolute size of particles shown here is not a key message, the relative fluorescence at a given size can be informative."

The referee points out that the work by Toprak and Schnaiter (2013) presented small soot particles that also exhibited relatively weaker fluorescence in FL1. This is consistent with the expectation that fluorescence intensity will scale strongly with particle size. Differences in particle size could also impact clustering separation properties somewhat, and so further investigation of clustering using multiple narrow size ranges of different types of particles could further explore this process. This exhaustive process was beyond the scope of this work, however.

To make sure these points are clear in the revised manuscript we have added the following text at L327:
"It is also important to note here that the method of aerosolization for each particle type plays an important role in the observed size distribution and so results involving laboratory particles should be interpreted with this in mind. Observed fluorescence properties, in contrast, are expected to be conserved at a given particle size and intrinsically related to particle composition."

[R1.7] L384: Would we expect to be able to differentiate between 2 different particles of the same type with such coarse spectral resolution?

[A1.7] The referee's implied point is correct. No, we would not expect to be able to separate between very similar types of particles using such coarse resolution as is available in the WIBS.

Frankly, the fact that HAC paired with WIBS data was able to separate as well as it did was somewhat remarkable and surprising. To make the point clearer, we added text at the end of that paragraph as follows at L390:
"…separating more finely to quantify differences between types of individual biological particles is ~~likely to be~~ significantly more challenging *and not likely to be possible in most situations.*"

[R1.8] L415: Again I wonder if the use of too coarsely separated bins may compromise the 9-sigma thresholding and cause misclassification?

[A1.8] This question also loops back to [R1.3] and stems from a miscommunication. Values of the five WIBS data parameters were not separately binned (either during the logging process or when used as recorded), but are input into the cluster algorithm in the same spacing provided in the raw output of the instrument. The bin resolution is therefore limited by the WIBS optics and PMT settings.

Further, fluorescence intensity is relayed by a integer units between 0 and 2047, and resolution is not a limiting factor. For example, see Figure 5 of the Savage et al. 2017 paper. Biological particles typically exhibit median fluorescence intensity much higher than non-biological particles, thus using different threshold strategies can help separate particle classes from one another by this strategy.

[R1.9] L514: Can the authors comment on the applicability of their findings to new high resolution UV-LIF instruments that are beginning to become commercially available. Some of these new instruments have significantly more channels/greater fluorescent resolution than the WIBS.

[A1.9] This is a helpful suggestion. To extend the applicability of results, the text was amended as follows:
"Results here are ~~only~~ generally extendable to other UV-LIF instruments, *however, whether they offer single or many channels of emission spectral resolution, in that the methods of particle pre-preparation and the impact of particle number ratio are likely to relay similar effects on clustering strategy."*

[R1.10] Technical corrections
L63: instruments, not instrument.
L370: grains, not gains.
L112: Suggest "Experimental and Computational Methods"
L131: "each of the three"
L181: "was the best"

[A1.10] All typos corrected.