Atmos. Meas. Tech., 6, 337–347, 2013
www.atmos-meas-tech.net/6/337/2013/
doi:10.5194/amt-6-337-2013

Atmospheric Measurement Techniques
Open Access

Cluster analysis of WIBS single-particle bioaerosol data

N. H. Robinson[1], J. D. Allan[1,2], J. A. Huffman[3], P. H. Kaye[4], V. E. Foot[5], and M. Gallagher[1]

[1]The Centre for Atmospheric Science, School of Earth Atmospheric and Environmental Science, The University of Manchester, Manchester, UK
[2]The National Centre for Atmospheric Science, The University of Manchester, Manchester, UK
[3]Department of Chemistry and Biochemistry, University of Denver, CO, USA
[4]Centre for Atmospheric & Instrumentation Research, STRI, University of Hertfordshire, Hatfield, AL10 9AB, UK
[5]DSTL, Porton Down, Salisbury, Wiltshire, SP4 0JQ, UK

Correspondence to: M. Gallagher (martin.gallagher@manchester.ac.uk)

Received: 24 July 2012 – Published in Atmos. Meas. Tech. Discuss.: 7 September 2012
Revised: 21 December 2012 – Accepted: 25 January 2013 – Published: 13 February 2013

Abstract. Hierarchical agglomerative cluster analysis was performed on single-particle multi-spatial data sets comprising optical diameter, asymmetry and three different fluorescence measurements, gathered using two dual Wideband Integrated Bioaerosol Sensors (WIBSs). The technique is demonstrated on measurements of various fluorescent and non-fluorescent polystyrene latex spheres (PSL) before being applied to two separate contemporaneous ambient WIBS data sets recorded in a forest site in Colorado, USA, as part of the BEACHON-RoMBAS project. Cluster analysis results between both data sets are consistent. Clusters are tentatively interpreted by comparison of concentration time series and cluster average measurement values to the published literature (of which there is a paucity) to represent the following: non-fluorescent accumulation mode aerosol; bacterial agglomerates; and fungal spores. To our knowledge, this is the first time cluster analysis has been applied to long-term online primary biological aerosol particle (PBAP) measurements. The novel application of this clustering technique provides a means for routinely reducing WIBS data to discrete concentration time series which are more easily interpretable, without the need for any a priori assumptions concerning the expected aerosol types. It can reduce the level of subjectivity compared to the more standard analysis approaches, which are typically performed by simple inspection of various ensemble data products. It also has the advantage of potentially resolving less populous or subtly different particle types. This technique is likely to become more robust in the future as fluorescence-based aerosol instrumentation measurement precision, dynamic range and the number of available metrics are improved.

1 Introduction

Primary biological aerosol particles (PBAPs) are those which are emitted or suspended directly from the biosphere to the atmosphere, and as such are composed of biological matter (Després et al., 2012). These aerosols can consist of the following: viruses (0.01–0.3 µm); bacteria and bacteria agglomerates (0.1–10 µm); fungal and plant spores (1–30 µm); and pollen (5–100 µm), as well as fragments thereof and of plant or animal matter (Després et al., 2012; Elbert et al., 2007). PBAPs can affect human health as allergens or through the transmission of disease, either naturally or through acts of bioterrorism (Cresti and Linskens, 2000). There is evidence that PBAPs may influence the hydrological cycle and climate by initiating warm ice nucleation processes (Christner et al., 2008; Möhler et al., 2007; Pratt et al., 2009; Prenni et al., 2009) or acting as giant cloud condensation nuclei (Möhler et al., 2007; Pope, 2010).

It is clear that the PBAP classification consists of aerosol from various diverse sources which may have wide reaching effects in the atmosphere. In order to predict these potential effects under future emissions scenarios, it is useful to be able to identify the group to which a measured PBAP belongs. To date, this has largely been achieved by the use of off-line techniques, which, whilst allowing accurate identification of different aerosols, are labour-intensive, have poor time resolution and introduce significant identification biases. Several light-induced fluorescence techniques have recently been developed which characterise the auto-fluorescence of particles, utilizing the presence of certain biofluorophores such as NAD(P)H, riboflavin, and tryptophan as indicators of PBAP

material (Hill et al., 2001; Huffman et al., 2010; Kaye et al., 2005; Pöhlker et al., 2012; Sivaprakasam et al., 2004, 2011; Pan et al., 2007, 2012; Pinnick et al., 2013).

Here we focus upon development of analysis techniques for the Wideband Integrated Bioaerosol Sensor (WIBS) range of auto-fluorescence detectors (Foot et al., 2008; Gabey et al., 2010, 2011; Kaye et al., 2005). We demonstrate the application of a cluster analysis technique to the WIBS single-particle data, allowing for robust statistical resolution of different PBAP subgroups.

2 The Wideband Integrated Bioaerosol Sensor

The measurements reported here were performed using two individual dual Wideband Integrated Bioaerosol Sensors (Foot et al., 2008; Gabey et al., 2010; Kaye et al., 2005; Stanley et al., 2011) – a model 3 (WIBS3) and a model 4 (WIBS4). In both these variants, the single-particle elastic scattering intensity (at 633 nm) is measured in the forward direction and at an angular range centred at 90°. These measurements are then used to infer the particle optical-equivalent diameter, D_O. The forward scattering component is measured by a quadrant photomultiplier tube that allows for measurement of the variation in azimuthal scattering from the particle. This in turn can be related to particle asymmetry or shape via an asymmetry factor, AF (e.g. Gabey et al., 2010). This sizing measurement triggers subsequent pulses from filtered xenon flash-lamps at 280 nm and 370 nm, designed to excite molecules such as tryptophan and nicotinamide adenine dinucleotide phosphate (NAD(P)H) respectively within the particle. Any resultant fluorescence is measured in two wavelength regimes named FL1 and FL2. This gives rise to three separate fluorescence channels: in FL1 and FL2 following the 280 nm excitation (named FL1_280 and FL2_280) and in FL2 following the 370 nm excitation (named FL2_370). The FL1 and FL2 fluorescence detection regimes overlap spectrally in the WIBS3 but have been separated in the WIBS4. There is no FL1_370 channel as the 370 nm light pulse lies within the FL1 detection regime, which leads to saturation. NAD(P)H does not fluoresce in the FL1 wavelength regime and riboflavin only weakly, while proteins and amino acids are more fluorescent in this channel. Table 1 details the fluorescence excitation and detection regimes for the two WIBS models. The WIBS4 also incorporates additional improvements to the optics configuration, excitation light delivery, sample inlet and logging software. A fluorescence baseline is determined from measurements of fluorescence when the xenon sources are fired in the absence of particles. This baseline has been subtracted from all fluorescence measurements presented here. In total, the WIBS provides five different measurements of each particle that are used in subsequent analyses herein: optical size, asymmetry factor, and three fluorescence measurements.

Previous work has identified different classes of PBAP using the physical properties measured by the WIBS instrument (Gabey et al., 2010, 2011; Gabey, 2011). However, this has so far been achieved by inspection of ensemble histograms which are then compared with particle standard measurements. This approach is labour-intensive, vulnerable to error, and may lead to the oversight of minor but important PBAP subgroups. It also does not easily lend itself to the production of concentration time series of PBAP subgroups necessary for more detailed understanding of particle emission sources.

Various cluster analysis techniques have previously been used to classify single-particle fluorescence data (Pinnick, 2004; Pan et al., 2007, 2012; Pinnick et al., 2013) and mass spectral data (Murphy et al., 2003), as well as back trajectories (Cox et al., 2005; Kalkstein et al., 1987; Robinson et al., 2011). In addition, neural networks have been trained to dynamically classify single-particle mass spectral data (Song et al., 1999). These studies have successfully demonstrated various approaches for objectively reducing large data sets so that they become easier to interpret, but have not yet been applied to data from WIBS or similar commercially available instruments. Previous studies have also focused on relatively short monitoring times (several days), in contrast to the data analysed here which cover several weeks. The following section identifies the most appropriate approach for the identification of a measured particle type. Firstly, several different approaches for identifying particle groups by analysing a subset of the data are discussed. This is followed by a discussion of particle attribution approaches, where the particles that were not included in the data subset are compared to and allocated to the previously identified groups. This allows for the construction of concentration time series of the different particle types for the entire measurement periods while only performing time-intensive calculations on representative subsets of the data.

3 Analysis techniques

The choice of particle grouping technique depends on the goals of the analysis and the properties of a given data set. We have chosen the following criteria as fundamental to suitable WIBS single-particle data analysis:

1. It should not require any assumptions about the types of particles present in the data set as this precludes the identification of PBAP types that have not previously been characterised using similar measurements.

2. It should not require any assumptions about relative group sizes, as different types of PBAP can be present in very different concentrations.

The technique also need not be dynamic, as WIBS analysis is performed offline. Neural network techniques have many attractive qualities such as their dynamic grouping, efficiency

Table 1. The excitation and detection wavelengths of the two WIBS models.

	FL1_280		FL2_280		FL2_370	
	Excitation	Detection	Excitation	Detection	Excitation	Detection
WIBS3	280 nm	320–600 nm	280 nm	410–600 nm	370 nm	410–600 nm
WIBS4	280 nm	310–400 nm	280 nm	420–650 nm	370 nm	420–650 nm

and accumulation of skill. However, they need prior training with measurements of different particle types, which requires assumptions about the types of particles present and so can lead to systematic misinterpretation. Cluster analysis is more suitable for WIBS data sets as it requires no such assumptions. The so-called k-means approach is a common, efficient cluster analysis technique. However, it tends to produce clusters of similar group size and spatial extent (Everitt, 1993), rendering it unsuitable for grouping PBAP. Hierarchical agglomerative (HA) cluster analysis meets all of the stated criteria (Everitt, 1993).

In HA cluster analysis each measured particle is initially considered to represent its own single-membered cluster. The algorithm identifies two clusters with the highest degree of similarity, which are then agglomerated into a new cluster. This step is repeated until all particles populate a single cluster. The analyst is then required to determine which step (number of clusters) most appropriately represents the data, which is a subjective process, but may be informed by several statistics. There are several different HA cluster analysis algorithms, each defined by the respective metric used for comparing the similarity of clusters.

The average-linkage HA cluster analysis algorithm is used herein as it is regarded as being robust and is conducive to groups of different size (Everitt, 1993; Kalkstein et al., 1987). It has the unique quality that it minimises the sum of squares within (SSW) cluster groups whilst maximising the sum of squares between (SSB) cluster groups. Average-linkage defines the two most similar clusters as those with the smallest distance across an n-dimensional space, where n is the number of measurements made of each particle (five in the case of the WIBS). The distance between two clusters is defined as the average squared Euclidian distance between all possible pairs of particles, one from each cluster, or

$$L_{\mathrm{A,B}} = \frac{1}{pq}\sum_{i=1}^{p}\sum_{j=1}^{q}\|\boldsymbol{A}_i - \boldsymbol{B}_j\|^2, \quad (1)$$

where $L_{\mathrm{A,B}}$ is the distance between clusters, $\boldsymbol{A}$ is the coordinate vector of cluster A which contains p members, and $\boldsymbol{B}$ is the coordinate vector of cluster B with q members. The use of Euclidian distances assumes symmetrically distributed data, so any variables that appear to be log-normally distributed are handled as their logarithms so as to give a more symmetric distribution. The data set is then z-score normalised before analysis.

The choice of the optimum number of clusters to retain is a subjective step, but it may be informed by various metrics (Everitt, 1993; Kalkstein et al., 1987). In average-linkage clustering the suitability of a solution of N clusters may be assessed by inspecting the coefficient of determination:

$$R^2 = 1 - \sum_{N}\frac{\text{sum of squares within groups}}{\text{total sum of squares}} \quad (2)$$

where a sharp decrease as N decreases is an indicator of the number of clusters to retain (Kalkstein et al., 1987; Robinson et al., 2011). An increase in the root mean squared (RMS) distance between clusters is an indication that two dissimilar clusters have been agglomerated (Cape et al., 2000). Additionally, the number of major clusters at each step is defined as being the number of clusters that are greater than half the mean cluster group size (Loureiro et al., 2004; Zoubi, 2009). This final metric is useful for assessing statistically insignificant clusters, but implicitly assumes that clusters are a similar size. There is no robust way of determining which clusters are insignificant (i.e. due to rogue measurements) and which clusters are significant (i.e. due to rare but important particle types). Any cluster deemed to be major by this metric should be retained in the subsequent analysis. Ultimately, due to the potential for radically different cluster group sizes, the analyst must decide which of the most minor clusters are unlikely to be representative of a physical particle type, and thus should be discarded. It should be noted that these statistics may indicate more than one solution is statistically significant. In such a case any indicated solution may be employed, with both being physically representative. If the solution comprises a greater number of clusters than there are particle types, then cluster time series will be split, and conversely if the solution comprises fewer clusters than there are particles types, then cluster time series will be conflated.

An average-linkage clustering algorithm was incorporated into the pre-existing suite of WIBS analysis tools, the WIBS AnalysiS Program (WASP). The routine was written using Igor Pro[1], with the numerical routines used to calculate cluster distances written in C and compiled as an external operation (XOP) library, in order to improve performance. A synthetic test data set was generated, consisting of three groups of two-dimensional points. Each group consisted of randomly generated points normally distributed around different

[1]WaveMetrics Inc., OR, USA.

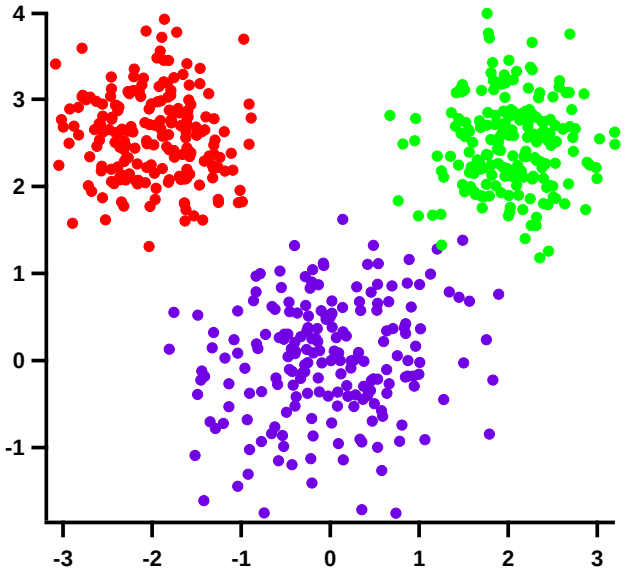

Fig. 1. Input to cluster analysis routine. Three synthetic, separately generated groups (differentiated by colour) of random, normally distributed data of arbitrary units centred around three separate points.

centres (Fig. 1). The WASP average-linkage routine statistics indicate that the three-cluster solution is optimum (Fig. 2). This solution correctly attributes 99 % of points to their original groups (Fig. 3). The only incorrect determinations are of points at the boundary between purple and green points.

4 Cluster analysis of WIBS data

The application of this approach to WIBS data presents some additional issues. Firstly, an implicit assumption of cluster analysis is that clustered particle types are static, that is that they do not evolve in the atmosphere through chemical or physical processing. When this is not the case, one particle type may be resolved as two or more clusters which represent different stages in the evolution of the particle. Additionally, the variables used in clustering should ideally not be interdependent, but, for any given particle composition, larger particles will fluoresce more intensely, despite no increase in their quantum yield (inherent ability to fluoresce). This means that WIBS fluorescence measurements are a convolution of particle size and fluorescence quantum yield. Inspection of WIBS measurements of monodisperse polystyrene latex spheres[2] (PSLs), which serve as particles of consistent inherent fluorescent ability but different sizes, shows that this effect is not compensated for by normalisation to the total elastic scattering or side scattering measurements also provided by the WIBS. As such, cluster analysis was performed using un-normalised WIBS fluorescence measurements. It

[2]Manufactured by Polysciences Inc., PA, USA.

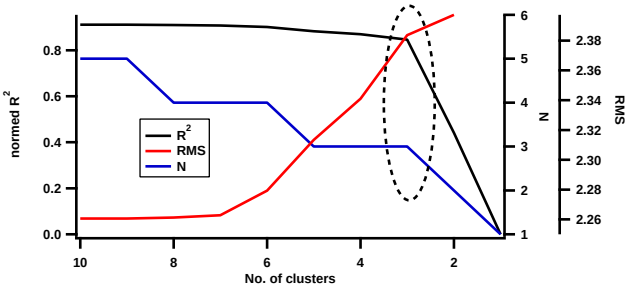

Fig. 2. Average-linkage statistics. The optimum solution as indicated by the statistics is highlighted.

should be noted that clustering will be weighted towards resolving particle groups that are separated by size, at the expense of resolving inherent fluorescent ability.

Additionally, a fluorescence detection channel can be saturated by some very large or very fluorescent particle, usually pollen or other large PBAPs. Typically around 5 % of ambient particles measured saturate at least one of the three fluorescence measurements. As the saturating particles are likely to be associated with a particular PBAP type, they have been included in the cluster analysis. During interpretation of the clustering solution, it should be noted that a cluster of saturating measurements may conflate different aerosol types (e.g. pollen subtypes) which would have been resolved had the detection range of the instrument been greater. Additionally, saturating aerosols may be conflated with highly fluorescent, but not saturating, aerosols, which can appear close in fluorescence space despite having relatively different quantum yields.

Data are assumed to be normally or log-normally distributed. In reality, the distribution of the data for a given measurement type is a convolution of measurement noise and the physical distribution of that property, with the relative contribution of each to the combined distribution related to their width. Inspection of the PSL data showed the inherent measurement noise of the WIBS to be normally distributed. Inspection of ambient data showed the overall distribution of size and AF measurements to be log-normally distributed, so these measurements were converted to log space prior to clustering. The distribution of fluorescence data is more complicated, with measurement values of zero and full saturation both possible. Given this, fluorescence measurements are assumed to be normally distributed. If this assumption is wrong, particles of low fluorescence are less likely to be resolved as separate clusters.

The computer processing time for the cluster analysis of a given data set grows approximately as the square of the size of the data set. An ambient data set may consist of measurements of $\sim 1 \times 10^6$ particles, which is impractically large for the WASP cluster analysis routine. Instead clusters are characterised using a randomly chosen subset of

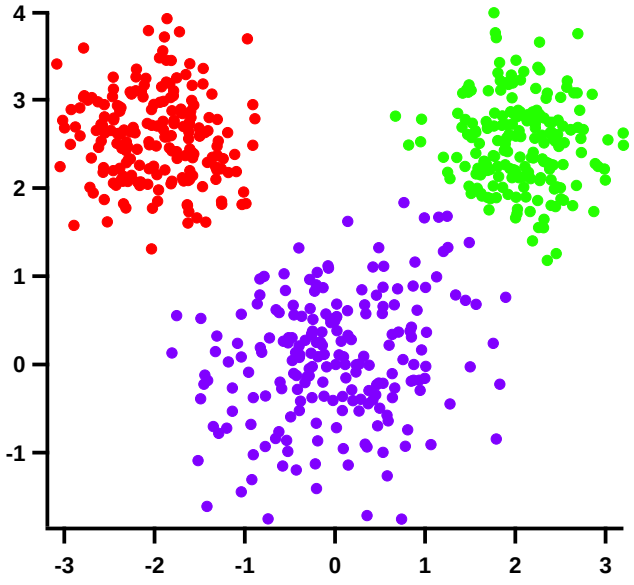

Fig. 3. Cluster analysis results, three-cluster solution. Each cluster indicated by colour. Three separate groups are resolved.

$\sim 1 \times 10^4$ particles, which takes approximately 4 h[3]. Once a suitable clustering solution has been chosen by inspection of the statistics, the remaining data are assigned to the different clusters by comparison to the cluster centroid. Measurements are again converted so that they are symmetrically distributed for this assignment. If the data belonging to each cluster form a distinct mode, then the mean and standard deviation are calculated from a Gaussian fit. This has the advantage of accurately identifying the modal centre, even if the entire distribution does not fall within the measurement range of the instrument. If a mode is not apparent (for instance when fluorescence measurements are saturated or zero, or there are a small number of measurements), the mean and standard deviation are calculated from the data themselves. Several different attribution algorithms were tested to find the most appropriate.

Two metrics can be used to assess the similarity of a particle measurement to a cluster. Firstly, the proximity of an individual measurement to the cluster centroid can be calculated after normalising each variable by its population standard deviation to account for differences in magnitude and variability. This is henceforth referred to as "population normalised distance" and is expressed by

$$d_i = \left| \frac{\boldsymbol{c}_i - \boldsymbol{p}}{\boldsymbol{\sigma}_{\text{pop}}} \right|, \tag{3}$$

where d_i is the distance of the particle measurement from cluster i, $\boldsymbol{c}_i$ is the position vector of cluster i in n-dimensional space, where n is the number of measured variables, $\boldsymbol{p}$ is the position vector of the particle in n-dimensional space, and $\boldsymbol{\sigma}_{\text{pop}}$ are the standard deviations of each measured variable across the entire data subset used in the cluster analysis. This approach does not take into account the spread (instrumental or physical) in the cluster distributions, but merely compares a particle measurement to the cluster modal centre.

Secondly, the population normalised distance approach can be extended by expressing the distance in each dimension in terms of the number of cluster standard deviations. This is henceforth referred to as "cluster normalised distance" and is expressed by

$$d_i = \left| \frac{\boldsymbol{c}_i - \boldsymbol{p}}{\boldsymbol{\sigma}_i} \right|, \tag{4}$$

where the symbols have their previous meaning and $\boldsymbol{\sigma}_i$ is a vector of the standard deviations of cluster i for each of the measured variables. This approach is conceptually pleasing in that it accounts for the spread of the variable values within the cluster and so represents the statistical uncertainty in apportioning a single-particle measurement to one cluster or another. However, this approach relies on the standard deviations of the distributions being precise. In practice, some clusters can display standard deviations that do not reflect the true spread of variable, which can then lead to systematic misattribution. This can be the case where less populous clusters do not form strong modes. It may also occur where standard deviations are estimated for modes that do not fall entirely within the measurement range of the instrument.

There are then two ways to use either of these d_i metrics to apportion the particle to a cluster. Firstly, the particle may be apportioned to the cluster which has the smallest d_i value, henceforth referred to as "simple attribution". Secondly, a fraction of each particle's count may be apportioned to each cluster that is inversely proportional to the distance of the particle from the cluster, such that the total of the fractions for any particle is unity. This is henceforth referred to as "fuzzy attribution", and the fraction attributed to cluster i is expressed by

$$F_i = \left(d_i \sum \frac{1}{d_i} \right)^{-1}, \tag{5}$$

where the symbols have their previous meanings. Any particles that are further away than a limit distance, d_l, are considered insignificant and deemed "unclassified". d_l is chosen as the minimum value, which also results in the unclassified particles being a minor group.

5 Cluster analysis of polystyrene latex spheres

Five different PSL types were measured sequentially using the WIBS4: 0.99 ± 0.01 µm standard[4]; $1 \pm < 0.1$ µm fluorescent; 1.90 ± 0.02 µm[5]; 3.005 ± 0.027 µm standard[4]; and

[3]Using a 3.4 GHz quad core processer, 8 GB RAM, 64-bit OS.

[4]Manufactured by Polysciences Inc., PA, USA.

[5]Manufactured by Duke Scientific Corp., CA, USA.

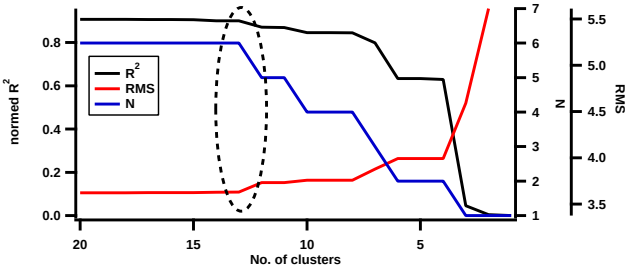

Fig. 4. PSL cluster analysis statistics. The 13-cluster solution was chosen due to the concomitant drop in R^2 and N, and the rise in RMS.

4.76 ± 0.04 µm standard[4]. The excitation maxima of the fluorescent PSLs are 365 nm, 388 nm and 412 nm, with respective emission maxima at 447 nm, 447 nm and 473 nm. As such, the fluorescent PSL would be expected to be detected mainly in the FL2_370 channel, with potential contributions to the other channels, depending on the width of the excitation/emission spectra. It should be noted that standard PSLs are also fluorescent but to a lesser extent, with fluorescence occurring due to 280 nm excitation. The modal values and relative standard deviations of the data input to the cluster analysis algorithm are shown in Table 2. The 13-cluster solution was chosen due to the observed decrease in R^2 and N, and the concomitant rise in RMS (Fig. 4). Of these 13 clusters, the six major clusters (as defined in Sect. 3) were retained for subsequent analysis (Table 3). It should be noted that, in this instance, the different PSLs are likely to be present in concentrations of a similar order of magnitude, meaning that each PSL type is likely to be resolved as a major cluster. Figure 5 shows the input single-particle size measurements as a function of time, coloured by cluster, and, below that, a comparison of cluster concentration time series generated using the attribution methods described above. Both fuzzy attribution sets used a significant distance limit (d_l) of five, which was set at the minimum value at which most particles are successfully attributed.

The PSL cluster analysis successfully resolves much of the data, with most of the PSL types individually represented by a cluster. In particular, the 3 and 4.76 µm PSL data are separately resolved as clusters E and F respectively. The 1 µm fluorescent PSL data are successfully resolved, although, in this solution, they are split between clusters A and B, which are qualitatively similar. It is not clear if this split is physically real (on the basis of different AF modes) or artificial. The 1 µm non-fluorescent PSL data are represented by cluster C. The majority of 2 µm PSL data are resolved as cluster D. However, a significant amount of these data belong to cluster C, apparently erroneously. This is likely due to the similarity of the 1 µm and 2 µm PSLs, which are very close in WIBS measurement space.

Table 2. Average modal centres of PSL measurements input to cluster analysis algorithm.

	1 µm	1 µm fl	2 µm	3 µm	4.76 µm
FL1_280	89 ± 0.2	14 ± 0.6	377 ± 0.1	860 ± 0.1	2083 ± 0.1
FL2_280	5 ± 1.8	2038 ± 0.1	10 ± 1.2	54 ± 3.7	252 ± 0.2
FL2_370	6 ± 2.2	1543 ± 0.4	9 ± 1.7	121 ± 0.4	241 ± 0.3
D_O (µm)	1.18 ± 1.3	1.19 ± 1.3	1.91 ± 1.2	3.49 ± 1.1	5.16 ± 1.1
AF	7.3 ± 1.4	7.1 ± 1.4	3.8 ± 1.6	4.7 ± 1.4	5.8 ± 1.4

The population normalised distance simple attribution approach appears to represent the data more satisfactorily than the other attribution algorithms. It generates concentration time series reflective of the clustering solution, with the exception of very small concentrations of cluster C particles during the introduction of 3 and 4.7 µm PSL. The two cluster normalised distance approaches attribute the majority of the non-fluorescent PSL to cluster C, presumably because cluster C has a greater spread in values than the other clusters. The population normalised distance fuzzy attribution approach, while correctly attributing the majority of particles to the correct clusters, attributes a significant number of the particles to other clusters.

The population normalised distance simple attribution is used for the rest of the presented analysis, given the combined advantages of transparent methodology, lack of sensitivity to potentially spurious distribution widths, and the lack of the need for setting a subjective distance limit.

6 Cluster analysis of two ambient WIBS data sets

The WIBS3 and WIBS4 were deployed as part of the Biohydro-atmosphere interactions of Energy, Aerosols, Carbon, H_2O, Organics and Nitrogen–Rocky Mountain Biogenic Aerosol Study project (BEACHON-RoMBAS[6]), which was performed between 20 June 2011 and 23 August 2011 in the Manitou Experimental Forest, 35 km west of Colorado Springs, CO, USA, 2300 m a.s.l. This project aims to investigate the effect biogenic aerosol emissions have on regional precipitation in the central US, and a full characterisation of aerosol properties and fluxes will be presented in Robinson et al. (2013). The WIBS3 was positioned around 200 m away from the main measurement site and sampled from ~ 1 m above the forest floor via ~ 0.5 m of $1/4''$ o.d. stainless steel tubing. The WIBS4 sampled via ~ 0.5 m of $1/4''$ stainless steel tubing as part of an automated profiling system running up the side of the main site measurement tower. WIBS4 profiles were regularly performed between 3.5 m and 20 m above the forest floor, measuring below, in, and above the forest canopy.

[6] http://web3.acd.ucar.edu/beachon/.

Table 3. Cluster average values and relative standard deviations of the six major clusters of the 13-cluster solution. Bottom row shows the number of constituent measurements. Minor clusters have been disregarded.

	A	B	C	D	E	F
FL1_280	12 ± 0.5	11 ± 0.7	229 ± 0.7	373 ± 0.1	889 ± 0.1	2106 ± 0.0
FL2_280	2060 ± 0.0	2060 ± 0.0	12 ± 4.6	9 ± 1.2	19 ± 1.6	212 ± 0.2
FL2_370	1859 ± 0.0	1853 ± 0.0	11 ± 5.3	7 ± 1.8	15 ± 1.5	112 ± 0.6
D_O (µm)	1.07 ± 1.2	1.10 ± 1.3	1.38 ± 1.5	1.82 ± 1.2	3.43 ± 1.1	5.02 ± 1.1
AF	3.9 ± 1.1	7.2 ± 1.3	5.0 ± 1.6	3.0 ± 1.3	4.46 ± 1.2	5.2 ± 1.3
#	281	1995	841	288	436	646

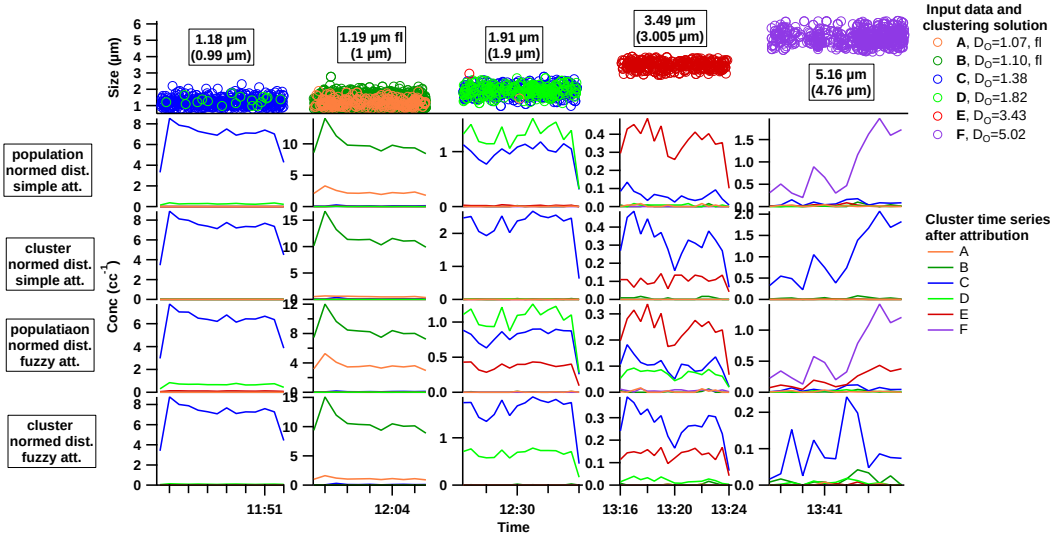

Fig. 5. Top plot shows the single-particle data input to the cluster analysis routine (~ 30 % of all the particles that were measured) coloured by their cluster, as defined in Table 3. Particle size and fluorescence for each retrieved cluster are detailed in the legend. Measured PSL optical diameters are detailed in boxes, with the actual PSL physical diameters in brackets. Below are cluster concentration time series after the remaining particles are attributed to the resolved clusters, using four different attribution algorithms, detailed on the left. The x-axis is discontinuous between the introduction of different PSLs. The y-axes are inconsistent between subplots.

For the WIBS3 data set, the four-cluster solution was selected based on the sharp drop in R^2 and rise in RMS (Fig. 6). All four clusters were retained for attribution (Table 4).

For the WIBS4 data set, the ten-cluster solution was selected based on the drop in R^2 and concomitant rise in RMS distance (Fig. 7). The six most populous of these clusters were retained for attribution, with the discarded clusters comprising five measurements or fewer (Table 5). Note that the three-cluster solution is also statistically significant, but it was considered likely that it was conflating particle types.

There is a paucity of published work characterising the measurement response of the WIBS to different aerosol types under controlled laboratory conditions. Tentative physical interpretations are presented here based on the existing literature; however, interpretation of the results of this clustering technique will be facilitated by further characterisation work.

The refinements made between WIBS models are likely to make certain quantitative comparisons of measurements impossible: the FL1 and FL2 measurement regimes have changed between instruments (Table 1), which could potentially lead to measurements of slightly different fluorescence properties of the same aerosol population; the reduced instrument noise in the WIBS4 should generally lead to smaller standard deviation values; and the improved fluorescence detection has led to lower fluorescence baselines, particularly in the FL2_280 channel. However, the clustering solutions of the two WIBS models are qualitatively similar. The clustering statistics indicate that cluster analysis has resolved a larger number of clusters using the WIBS4 data set than using the WIBS3. This may be due to the greater precision of the WIBS4 allowing the resolution of particle groups that are conflated in analysis of the WIBS3 data. Inspection of the WIBS4 clustering solutions shows that clusters A_4 and B_4 are agglomerated in the nine-cluster solution, and clusters C_4 and D_4 are agglomerated in the six-cluster solution (with intervening solutions agglomerating discarded clusters). This

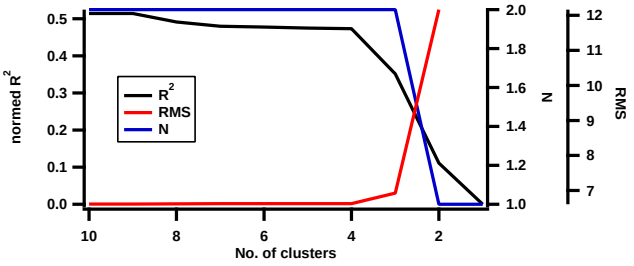

Fig. 6. Ambient WIBS3 data set clustering statistics. Four-cluster solution chosen due concomitant drop in R^2 and rise in RMS.

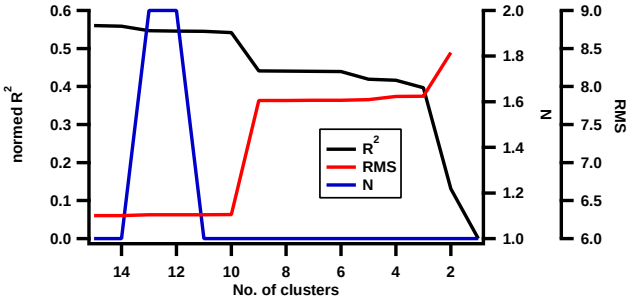

Fig. 7. Ambient WIBS4 data set clustering statistics. Ten-cluster solution chosen due concomitant drop in R^2 and rise in RMS.

Table 4. Cluster averages and relative standard deviations for the WIBS3 data set. Bottom row shows the number of constituent measurements.

	A_3	B_3	C_3	D_3
FL1_280	25 ± 3.1	1725 ± 6.6	87 ± 1.6	1542 ± 0.5
FL2_280	44 ± 1.4	230 ± 0.5	331 ± 0.4	1475 ± 0.2
FL2_370	90 ± 1.6	136 ± 0.1	1224 ± 0.3	1885 ± 0.1
D_O (µm)	1.6 ± 1.6	2.9 ± 1.6	3.1 ± 1.7	4.4 ± 1.6
AF	15.9 ± 1.8	21.2 ± 1.5	17.5 ± 2.2	18.5 ± 1.5
#	9670	456	243	43

suggests that these clusters are the most statistically similar of the six retained clusters, and, as such, their concentrations may need to be summed for comparison to the WIBS3 solution.

Clusters A_3, A_4 and B_4 are likely to represent the tail end of the ambient accumulation mode, being relatively abundant, small in diameter and non-fluorescent. As such, it is likely to comprise several different non-PBAP sources.

Clusters B_3, C_4 and D_4 show high fluorescence in FL1_280. The bacteria *P. syringae* and *P. fluorescens* have previously been shown to fluoresce strongly in FL1_280 using the WIBS3 (Gabey, 2011). Bacteria are often present in the atmosphere as bacteria aggregate clumps or as a constituent part of some other aerosol (Després et al., 2012). Aerosols containing culturable bacteria have been reported to have aerodynamic diameters of ~ 4 µm at several continental sites (Després et al., 2012; Tong and Lighthart, 2000; Wang et al., 2007), which is similar to the cluster diameters of ~ 2–4 µm. The relatively high cluster AFs are consistent with bacterial aggregates, which are expected to be highly asymmetric. Thus, this literature as it stands suggests these clusters may represent bacterial aggregates or some other bacteria-containing aerosol.

Clusters C_3 and E_4 both show high fluorescence in FL2_370, which has previously been found to be characteristic of grass smut fungal spores such as Bermuda grass smut and Johnson grass smut (Gabey, 2011) using the WIBS3. However, those species tend to be larger in size than the D_O of ~ 3 µm (6–8 µm and 6–10 µm respectively). It is possible that these clusters represent some other fungal spore. C_3 and E_4 have substantially different FL2_280 measurements, which is likely to be due to the stated differences between WIBS models. Previous work has not yet established if high FL2_280 levels are typical of grass smut fungal spore measurements in the WIBS4.

Clusters D_3 and F_4 are very likely to represent fungal spores, as they are highly fluorescent in all three channels, are relatively large and asymmetric (Gabey, 2011). While the average diameter is much smaller than that of pollen, it is likely that any pollen detected has been conflated with this cluster, as it is also highly fluorescent in all channels.

Population normalised distance simple attribution was used to generate concentration time series for the clusters resolved by each instrument. The time series gradient of scatter and Pearson's r values are shown in Table 6 with the time series for each cluster shown in Fig. 8. The time series from each instrument compare very well, especially considering that some of the less populous clusters are close to the limit of detection of the instruments. The time series show clear separation of different factors. The accumulation mode and smut fungal spore clusters were found to respond to meteorological variables such as precipitation and relative humidity (Huffman et al., 2013), while the bacteria and other fungal spore clusters show a strong nocturnal profile.

The diurnal profiles are largely consistent with the physical interpretation of the clusters. It should be noted that the extreme upper size range of the accumulation mode, which these clusters represent, may not have the same source profile as the rest of the accumulation mode. The nocturnal elevation of the concentrations of the other fungal spore clusters is consistent with the literature, which reports some kinds of active fungal spore emissions at night when the relative humidity is high (Després et al., 2012; Elbert et al., 2007; Gabey et al., 2010). The nocturnal elevation of the bacteria clusters is inconsistent with previous bacteria measurements, which tend to show peak culturable bacteria concentrations during the day (Shaffer and Lighthart, 1997). However, data from comparable sites are limited. It is also possible that the WIBS technique is more sensitive to non-culturable bacteria that

Table 5. Cluster averages and standard deviations for the WIBS4 data set. Bottom row shows the number of constituent measurements.

	A_4	B_4	C_4	D_4	E_4	F_4
FL1_280	5 ± 3.8	30 ± 2.1	2087 ± 0.0	1124 ± 0.6	86 ± 1.5	2110 ± 0.0
FL2_280	98 ± 1.4	702 ± 0.5	1486 ± 0.3	518 ± 0.5	1849 ± 0.2	2055 ± 0.0
FL2_370	80 ± 1.3	620 ± 0.5	492 ± 0.6	119 ± 0.9	1893 ± 0.1	1822 ± 0.1
D_O (µm)	1.6 ± 1.6	2.1 ± 2.0	3.5 ± 1.4	2.4 ± 1.5	2.8 ± 1.8	4.9 ± 1.4
AF	8.6 ± 2.0	9.5 ± 2.3	20.6 ± 1.8	15.6 ± 1.9	12.3 ± 3.5	26.8 ± 1.8
#	7934	384	138	92	91	27

Table 6. Gradient, m, of straight line least squares regression fit through zero of scattered data, and Pearson's r, for cluster time series from each WIBS instrument. Only profile data from below 5 m were used to aid comparison.

WIBS3 vs. WIBS4	A_3, $A_4 + B_4$	B_3, $C_4 + D_4$	C_3, E_4	D_3, F_4
m	0.92	1.23	2.18	1.43
r	0.84	0.81	0.73	0.73

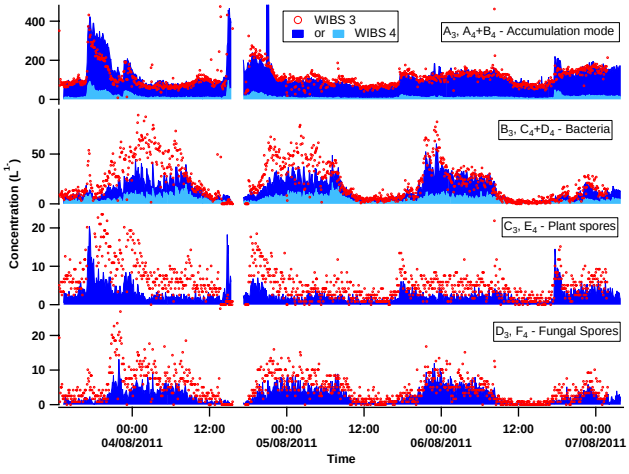

Fig. 8. Comparison of cluster time series over an exemplary period. WIBS3 time series shown by red points and WIBS4 time series shown by blue bars. Cluster names and their physical interpretation detailed in labels. A_4 (dark blue) and B_4 (light blue) are stacked for comparison to A_3, and C_4 (dark blue) and D_4 (light blue) are stacked for comparison to B_3. WIBS4 measurements from all profile heights are displayed.

are missed by off-line techniques, or insensitive to smaller bacteria aerosols which are detected on filters. It is also possible that these clusters represent some non-bacteria aerosol type which has yet to be characterised using the WIBS. The nocturnal increase seen in the other fungal spores and bacteria clusters may also be due to the collapse of the nocturnal boundary layer if sources are local . A full interpretation of the time series of these clusters, plus cluster gradient flux estimates, will be presented in Robinson et al. (2013).

Fluorescence scanning electron microscope (SEM) and DNA analysis of filter samples were performed as part of the same project (Huffman et al., 2013; Prenni et al., 2013). That analysis is consistent with the interpretation of the cluster analysis presented here, with identified species including Proteobacteria, Actinobacteria, Firmicutes, Bacteroidetes, Enterobacteriaceae and Pseudomonadaceae bacteria; Basidiomycota (club fungi) and Ascomycota (sac fungi) fungal spores; and smut fungal spores.

7 Conclusions

Hierarchical agglomerative cluster analysis was successfully applied to a subset of WIBS measurements. The remaining measurements were then attributed to the resolved clusters, allowing the generation of respective concentration time series. The approach was tested and verified on a controlled data set of PSL measurements. Several attribution approaches were compared, with the most effective being association of each particle with the cluster to which it is closest to in n-dimensional measurement space when normalised for variability and magnitude. The cluster analysis of PSL data, whilst it partially conflated two similar PSLs, successfully resolved most PSL groups.

The technique was then applied to two separate contemporaneous ambient WIBS data sets. To our knowledge, this is the first time cluster analysis has been applied to a data set of long-term online PBAP measurements. The average measurement values of clusters were qualitatively similar between the two instruments, if differences in instrument design are taken into account. The cluster concentration time series compare quantitatively well between the two instruments. The ambient cluster results were associated with aerosol types by comparison of the cluster measurement averages and time series to the existing literature. It appears that the cluster analysis resolved the following: accumulation mode aerosol; bacterial clusters; fungal smut spores; and other fungal spores. It should be noted that there is a paucity of work characterising the response of the WIBS to different PBAP types, and, as such, the physical interpretation presented here is tentative. Future studies should aim to present systematic laboratory characterisation of PBAP subtypes in

order to allow more rigorous interpretation of WIBS cluster analyses. Future WIBS models are expected to increase instrument precision and introduce more fluorescence measurement channels, which will vastly improve the effectiveness of this cluster analysis approach. Future work should aim to extend this method to real-time discriminatory PBAP monitoring.

Acknowledgements. Thanks to J. Crosier of the Centre for Atmospheric Science, the University of Manchester, for help making the Igor XOP. Thanks to Jose-Luis Jimenez, and Douglas Day of the University of Colorado, Boulder, for organisation of the site logistics. J. A. Huffman acknowledges University of Denver internal faculty funding for support. N. H. Robinson was supported by UK NERC grant NE/H019049/1.

Edited by: D. Toohey

References

Cape, J. N., Methven, J., and Hudson, L. E.: The use of trajectory cluster analysis to interpret trace gas measurements at Mace Head, Ireland, Atmos. Environ., 34, 3651–3663, doi:10.1016/S1352-2310(00)00098-4, 2000.

Christner, B. C., Morris, C. E., Foreman, C. M., Cai, R., and Sands, D. C.: Ubiquity of biological ice nucleators in snowfall, Science, 319, 1214, doi:10.1126/science.1149757, 2008.

Cox, M. L., Sturrock, G. A., Fraser, P. J., Siems, S. T., and Krummel, P. B.: Identification of regional sources of methyl bromide and methyl iodide from AGAGE observations at Cape Grim, Tasmania, J. Atmos. Chem., 50, 59–77, doi:10.1007/s10874-005-2434-5, 2005.

Cresti, M. and Linskens, H. F.: Pollen-allergy as an ecological phenomenon: a review, Plant Biosyst., 134, 341–352, doi:10.1080/11263500012331350495, 2000.

Després, V. R., Huffman, J. A., Burrows, S. M., Hoose, C., Safatov, A. S., Buryak, G., Fröhlich-Nowoisky, J., Elbert, W., Andreae, M. O., Pöschl, U., and Jaenicke, R.: Primary biological aerosol particles in the atmosphere: a review, Tellus B, 64, 15598, doi:10.3402/tellusb.v64i0.15598, 2012.

Elbert, W., Taylor, P. E., Andreae, M. O., and Pöschl, U.: Contribution of fungi to primary biogenic aerosols in the atmosphere: wet and dry discharged spores, carbohydrates, and inorganic ions, Atmos. Chem. Phys., 7, 4569–4588, doi:10.5194/acp-7-4569-2007, 2007.

Everitt, B.: Cluster Analysis, 3rd Edn., John Wiley & Sons, New York, 1993.

Foot, V. E., Kaye, P. H., Stanley, W. R., Barrington, S. J., Gallagher, M., and Gabey, A.: Low-cost real-time multiparameter bio-aerosol sensors, in: Proceedings of SPIE, 71160, 71160I–71160I–12, doi:10.1117/12.800226, 2008.

Gabey, A. M.: Laboratory and field characterisation of fluorescent and primary biological aerosol particles, Ph.D., University of Manchester, 2011.

Gabey, A. M., Gallagher, M. W., Whitehead, J., Dorsey, J. R., Kaye, P. H., and Stanley, W. R.: Measurements and comparison of primary biological aerosol above and below a tropical forest canopy using a dual channel fluorescence spectrometer, Atmos. Chem. Phys., 10, 4453–4466, doi:10.5194/acp-10-4453-2010, 2010.

Gabey, A. M., Stanley, W. R., Gallagher, M. W., and Kaye, P. H.: The fluorescence properties of aerosol larger than 0.8 µm in urban and tropical rainforest locations, Atmos. Chem. Phys., 11, 5491–5504, doi:10.5194/acp-11-5491-2011, 2011.

Hill, S. C., Pinnick, R. G., Niles, S., Fell, N. F., Pan, Y.-L., Bottiger, J., Bronk, B. V., Holler, S., and Chang, R. K.: Fluorescence from airborne microparticles: dependence on size, concentration of fluorophores, and illumination intensity, Appl. Optics, 40, 3005, doi:10.1364/AO.40.003005, 2001.

Huffman, J. A., Treutlein, B., and Pöschl, U.: Fluorescent biological aerosol particle concentrations and size distributions measured with an Ultraviolet Aerodynamic Particle Sizer (UV-APS) in Central Europe, Atmos. Chem. Phys., 10, 3215–3233, doi:10.5194/acp-10-3215-2010, 2010.

Huffman, J. A., Pöhlker, C., Prenni, A. J., DeMott, P. J., Mason, R. H., Robinson, N. H., Fröhlich-Nowoisky, J., Tobo, Y., Després, V. R., Garcia, E., Gochis, D. J., Harris, E., Müller-Germann, I., Ruzene, C., Schmer, B., Sinha, B., Day, D. A., Andreae, M. O., Jimenez, J. L., Gallagher, M., Kreidenweis, S. M., Bertram, A. K., and Pöschl, U.: High concentrations of biological aerosol particles and ice nuclei during and after rain, Atmos. Chem. Phys. Discuss., 13, 1767–1793, doi:10.5194/acpd-13-1767-2013, 2013.

Kalkstein, L. S., Tan, G., and Skindlov, J. A.: An Evaluation of three clustering procedures for use in synoptic climatological classification, J. Appl. Meteorol., 26, 717–730, 1987.

Kaye, P. H., Stanley, W. R., Hirst, E., Foot, E. V., Baxter, K. L., and Barrington, S. J.: Single particle multichannel bio-aerosol fluorescence sensor, Opt. Express, 13, 3583, doi:10.1364/OPEX.13.003583, 2005.

Loureiro, A., Torgo, L., and Soares, C.: Outlier detection using clustering methods: a data cleaning application, in: Proceedings of the Data Mining for Business Workshop, edited by: Soares, C., Moniz, L., and Duarte, C., Citeseer, 57–62, available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.61.7266&rep=rep1&type=pdf, 2004.

Möhler, O., DeMott, P. J., Vali, G., and Levin, Z.: Microbiology and atmospheric processes: the role of biological particles in cloud physics, Biogeosciences, 4, 1059–1071, doi:10.5194/bg-4-1059-2007, 2007.

Murphy, D. M., Middlebrook, A. M., and Warshawsky, M.: Cluster analysis of data from the particle analysis by Laser Mass Spectrometry (PALMS) instrument, Aerosol Sci. Tech., 37, 382–391, doi:10.1080/02786820300971, 2003.

Pan, Y.-L., Pinnick, R. G., Hill, S. C., Rosen, J. M., and Chang, R. K.: Single-particle laser-induced-fluorescence spectra of biological and other organic-carbon aerosols in the atmosphere: Measurements at New Haven, Connecticut, and Las Cruces, New Mexico, J. Geophys. Res., 112, D24S19, doi:10.1029/2007JD008741, 2007.

Pan, Y. L., Huang, H., and Chang, R. K.: Clustered and integrated fluorescence spectra from single atmospheric aerosol particles excited by a 263- and 351-nm laser at New Haven, CT, and Adelphi, MD, J. Quant. Spectrosc. Ra., 113, 2213–2221, doi:10.1016/j.jqsrt.2012.07.028, 2012.

Pinnick, R.: Fluorescence spectra of atmospheric aerosol at Adelphi, Maryland, USA: measurement and classification of single

particles containing organic carbon, Atmos. Environ., 38, 1657–1672, doi:10.1016/j.atmosenv.2003.11.017, 2004.

Pinnick, R. G., Fernandez, E., Rosen, J. M., Hill, S. C., Wang, Y., and Pan, Y. L.: Fluorescence spectra and elastic scattering characteristics of atmospheric aerosol in Las Cruces, New Mexico, USA: Variability of concentrations and possible constituents and sources of particles in various spectral clusters, Atmos. Environ., 65, 195–204, doi:10.1016/j.atmosenv.2012.09.020, 2013.

Pöhlker, C., Huffman, J. A., and Pöschl, U.: Autofluorescence of atmospheric bioaerosols – fluorescent biomolecules and potential interferences, Atmos. Meas. Tech., 5, 37–71, doi:10.5194/amt-5-37-2012, 2012.

Pope, F. D.: Pollen grains are efficient cloud condensation nuclei, Environ. Res. Lett., 5, 044015, doi:10.1088/1748-9326/5/4/044015, 2010.

Pratt, K. A., DeMott, P. J., French, J. R., Wang, Z., Westphal, D. L., Heymsfield, A. J., Twohy, C. H., Prenni, A. J., and Prather, K. A.: In situ detection of biological particles in cloud ice-crystals, Nat. Geosci., 2, 398–401, doi:10.1038/ngeo521, 2009.

Prenni, A. J., Petters, M. D., Kreidenweis, S. M., Heald, C. L., Martin, S. T., Artaxo, P., Garland, R. M., Wollny, A. G., and Pöschl, U.: Relative roles of biogenic emissions and Saharan dust as ice nuclei in the Amazon Basin, Nat. Geosci., 2, 402–405, doi:10.1038/ngeo517, 2009.

Prenni, A. J., Tobo, Y., Garcia, E., DeMott, P. J., McCluskey, C. S., Kreidenweis, S. M., Prenni, J. E., Huffman, J. A., Pöhlker, C., and Pöschl, U.: The impact of rain on ice nuclei populations at a forested site in Colorado, Geophys. Res. Lett., online first, doi:10.1029/2012GL053953, 2013.

Robinson, N. H., Newton, H. M., Allan, J. D., Irwin, M., Hamilton, J. F., Flynn, M., Bower, K. N., Williams, P. I., Mills, G., Reeves, C. E., McFiggans, G., and Coe, H.: Source attribution of Bornean air masses by back trajectory analysis during the OP3 project, Atmos. Chem. Phys., 11, 9605–9630, doi:10.5194/acp-11-9605-2011, 2011.

Robinson, N. H., Flynn, M. J., Foot, E. V., and Gallagher, M. W.: Biogenic aerosol charaterisation and fluxes in a North American boreal forest, Atmos. Chem. Phys. Discuss., in preparation, 2013.

Shaffer, B. and Lighthart, B.: Survey of culturable airborne bacteria at four diverse locations in oregon: urban, rural, forest, and coastal, Microb. Ecol., 34, 167–177, doi:10.1007/s002489900046, 1997.

Sivaprakasam, V., Huston, A. L.,Scotto, C., Eversole, J. D.:Multiple UV wavelength excitation and fluorescence of bioaerosols, Opt. Express, 12, doi:10.1364/OPEX.12.004457, 4457–4466, 2004.

Sivaprakasam, V., Lin H. B., Huston, A. L., and Eversole, J. D.:Spectral characterization of biological aerosol particles using two-wavelength excited laser-induced fluorescence and elastic scattering measurements, Opt. Express, 19, doi:10.1364/OE.19.006191, 6191–6208, 2011.

Song, X.-H., Hopke, P. K., Fergenson, D. P., and Prather, K. A.: Classification of single particles analyzed by ATOFMS using an artificial neural network, ART-2A, Anal. Chem., 71, 860–865, doi:10.1021/ac9809682, 1999.

Stanley, W. R., Kaye, P. H., Foot, V. E., Barrington, S. J., Gallagher, M., and Gabey, A.: Continuous bioaerosol monitoring in a tropical environment using a UV fluorescence particle spectrometer, Atmos. Sci. Lett., 12, 195–199, doi:10.1002/asl.310, 2011.

Tong, Y. and Lighthart, B.: The annual bacterial particle concentration and size distribution in the ambient atmosphere in a rural area of the Willamette Valley, Oregon, Aerosol Sci. Tech., 32, 393–403, doi:10.1080/027868200303533, 2000.

Wang, C.-C., Fang, G.-C., and Lee, L.: Bioaerosols study in central Taiwan during summer season, Toxicol. Ind. Health, 23, 133–139, doi:10.1177/0748233707078741, 2007.

Zoubi, B. A.: An effective clustering-based approach for outlier detection, Eur. J. Sci. Res., 28, 310–316, 2009.