Atmos. Meas. Tech., 8, 4979–4991, 2015
www.atmos-meas-tech.net/8/4979/2015/
doi:10.5194/amt-8-4979-2015

Atmospheric Measurement Techniques
Open Access
EGU

Evaluation of hierarchical agglomerative cluster analysis methods for discrimination of primary biological aerosol

I. Crawford[1], **S. Ruske**[1], **D. O. Topping**[1,2], **and M. W. Gallagher**[1]

[1]Centre for Atmospheric Science, SEAES, University of Manchester, Manchester, UK
[2]NCAS, National Centre for Atmospheric Science, University of Manchester, Manchester, UK

Correspondence to: I. Crawford (i.crawford@manchester.ac.uk)

Received: 13 May 2015 – Published in Atmos. Meas. Tech. Discuss.: 16 July 2015
Revised: 6 November 2015 – Accepted: 10 November 2015 – Published: 27 November 2015

Abstract. In this paper we present improved methods for discriminating and quantifying primary biological aerosol particles (PBAPs) by applying hierarchical agglomerative cluster analysis to multi-parameter ultraviolet-light-induced fluorescence (UV-LIF) spectrometer data. The methods employed in this study can be applied to data sets in excess of 1×10^6 points on a desktop computer, allowing for each fluorescent particle in a data set to be explicitly clustered. This reduces the potential for misattribution found in subsampling and comparative attribution methods used in previous approaches, improving our capacity to discriminate and quantify PBAP meta-classes. We evaluate the performance of several hierarchical agglomerative cluster analysis linkages and data normalisation methods using laboratory samples of known particle types and an ambient data set.

Fluorescent and non-fluorescent polystyrene latex spheres were sampled with a Wideband Integrated Bioaerosol Spectrometer (WIBS-4) where the optical size, asymmetry factor and fluorescent measurements were used as inputs to the analysis package. It was found that the Ward linkage with z-score or range normalisation performed best, correctly attributing 98 and 98.1 % of the data points respectively. The best-performing methods were applied to the BEACHON-RoMBAS (Bio–hydro–atmosphere interactions of Energy, Aerosols, Carbon, H2O, Organics and Nitrogen–Rocky Mountain Biogenic Aerosol Study) ambient data set, where it was found that the z-score and range normalisation methods yield similar results, with each method producing clusters representative of fungal spores and bacterial aerosol, consistent with previous results. The z-score result was compared to clusters generated with previous approaches (WIBS AnalysiS Program, WASP) where we observe that the subsampling and comparative attribution method employed by WASP results in the overestimation of the fungal spore concentration by a factor of 1.5 and the underestimation of bacterial aerosol concentration by a factor of 5. We suggest that this likely due to errors arising from misattribution due to poor centroid definition and failure to assign particles to a cluster as a result of the subsampling and comparative attribution method employed by WASP. The methods used here allow for the entire fluorescent population of particles to be analysed, yielding an explicit cluster attribution for each particle and improving cluster centroid definition and our capacity to discriminate and quantify PBAP meta-classes compared to previous approaches.

1 Introduction

Microorganisms influence climate through their physical and chemical interactions with the atmosphere. Recently there has been renewed interest in how primary biological aerosol particles (PBAPs) interact with and modify clouds. It has been shown that bacterial aerosol such as *Pseudomonas syringae* can act as ice nuclei (IN) at relatively warm temperatures (Möhler et al., 2007), which even in low concentrations can cause rapid cloud glaciation via the Hallet–Mossop process, leading to premature precipitation (Crawford et al., 2012).

It has been hypothesised that a feedback cycle exists where PBAPs associated with plants influence the formation and modification of clouds through ice formation to induce precipitation, creating an environment which is beneficial for plant and microbial growth and thus stimulating fur-

ther PBAP emission (Sands et al., 1982) – this is known as the bioprecipitation hypotheses, and potential links between long-term regional climatology and PBAP emissions have recently been suggested (Morris et al., 2014). One of the key drivers for new research into bioprecipitation is a need for more accurate quantification of cloud evolution and precipitation in weather and climate models given its potential impact.

Bioaerosols are now being included as important components in global climate models (Heald and Spracklen, 2009; Jacobson and Streets, 2009). Recently bioaerosol emission models were tested on European regional scales (Hummel et al., 2014) using real-time Wideband Integrated Bioaerosol Spectrometer (WIBS-4) data collected at rural and semi-rural sites in Germany and Finland (Toprak and Schnaiter, 2013; Schumacher et al., 2013). Validation of these models is reliant on a very limited number of studies, and the authors highlight the difficulty of applying such models to e.g. urban environments and cite the general paucity of high-resolution atmospheric PBAP data to constrain model results. Providing such data is paramount to improving model predictions and accurately assessing the impact of PBAP emissions on environment and health. Retrieving such data is reliant upon the applicability of detection methods described in the following section.

The focus of this study is to evaluate hierarchical agglomerative cluster analysis methods applied to WIBS ultraviolet-light-induced fluorescence (UV-LIF) data sets for the discrimination of primary biological aerosol. In this paper we describe the detection method and data preparation procedures before evaluating the performance of several common hierarchical agglomerative cluster analysis linkages and data normalisation methods using laboratory and ambient data sets.

Detection methods

The detection, classification and quantification of PBAPs remain a significant multidisciplinary technical challenge. Conventional techniques can be split into culturing and non-culturing techniques, both of which require the collection of particles onto a medium for offline analysis. Culturing techniques collect particles of interest onto a growth medium which is incubated for hours to days. The grown colonies are then counted microscopically, providing species identification but not quantification of their atmospheric concentration, making the technique unsuitable for estimating PBAP emissions (Gabey, 2011). Non-culturing techniques collect particles onto filters or in a liquid suspension, which is more suitable for estimating atmospheric concentrations but is not typically used for classification (Douwes et al., 2003). The major limiting factors of non-culturing techniques are that they are labour intensive, require long sampling periods and suffer from impactor sampling artefacts (e.g. particle fragmentation, obscuration), leading to erroneous enumeration. This makes it difficult to study emissions at the process level as some PBAPs, such as fungal spores and bacteria, display large diurnal variations with significant short-term episodic emissions, which would require an impractical number of samples to capture reliably. PBAPs including bacteria can undergo substantial instantaneous spikes in emissions compared to their baseline state in response to rainfall (Crawford et al., 2014; Hummel et al., 2014). These rapid emissions are important not only to capture peak concentrations but also to derive emission factors accurately and understand the underlying mechanisms.

UV-LIF spectrometers have become available which show early promise of classifying and quantifying bioaerosols by broad taxonomic class on a single-particle basis (Crawford et al., 2014; Gabey et al., 2013). This instrument is based on technology developed by the University of Hertfordshire Centre for Atmospheric and Instrumentation Research (CAIR). A full technical description of the WIBS instrument is given later in this manuscript. UV-LIF spectrometers work on the principle that PBAPs contain biofluorophores such as NAD(P)H, riboflavin, and tryptophan which auto-fluoresce when excited with UV radiation with the excitation, and detection bands of the WIBS are optimised to detect these common biofluorophores (Kaye et al., 2005). The single-particle, online nature of the technique yields far superior time resolution to the offline techniques discussed earlier, making it ideally suited to measuring PBAPs in a rapidly changing environment. The time resolution is limited by the counting statistics, with typically 1–5 min integration periods providing adequate sensitivity depending on ambient concentrations. This allows for better measurements of PBAP fluxes, which would be difficult using traditional offline methods.

Whilst UV-LIF spectrometers offer many advantages over traditional methods, discriminating between different bioaerosol classes and possible, non-biological fluorescent interferents remains an ongoing area of research (Toprak and Schnaiter, 2013). At present, UV-LIF spectrometers lack a common absolute reference standard, making comparison of measurements made between instruments difficult. Furthermore the lack of a calibration standard has impeded attempts to characterise PBAPs of interest which would greatly simplify classification by the utilisation of supervised learning techniques. In lieu of an absolute calibration method other techniques must be used to segregate particles by type when interpreting uncalibrated data sets.

2 WIBS UV-LIF instrumentation

A full technical description of the original WIBS measurement principles and its development is given by Kaye et al. (2005), Foot et al. (2008), Gabey et al. (2011) and Stanley et al. (2011). In the versions of the instrument used here ambient air is sampled at 2.38 L min^{-1}, with 10 % of the to-

tal as aerosol flow drawn through a 1.2 mm (inner diameter) tube to generate a single in-line aerosol beam intersecting a well-defined optical sensing region. The remainder of the flow is filtered and used as a sheath flow to stabilise the aerosol beam and minimise possible detrainment contamination of the optical surfaces within the scattering chamber. Single particles passing through the sensing region intercept a 635 nm diode laser beam, and the elastically scattered forward and sideways intensity is measured. A lookup table based on a standard Mie scattering model (Kaye et al., 2005) is used to convert the forward-scatter / side-scatter intensity ratio to optical diameter based on the instrument's response to NIST calibration polystyrene latex (PSL) spheres. The WIBS utilises a quadrant detector to measure the scattered intensity. The signal from each component quadrant is used to calculate an "average" optical diameter over the four scattering solid angles. In addition the standard deviation between the four signal intensities is used to provide a particle asymmetry factor (AF) as a proxy of particle morphology. AF is reported in arbitrary units (a.u.) and is based on measurements with calibration particles with different aspect ratios; corn starch flour was used to represent irregular particles, and ellipsoidal haematite particles were used as an analogue for rod-like bacterial particles as described in Kaye et al. (2007). AF ranges from 8 to 10 for nearly spherical particles and 20–100 for a rod- or fibre-like particles. The detectable particle "average optical diameter" range for WIBS-4 is $0.5 < D_o < 20\,\mu m$, with a 50 % detection at $D_{p50} = 0.8\,\mu m$ (Gabey et al., 2011). The WIBS size range is optimised to sample most airborne bacteria and fungal spores, but only very small pollen. Following initial particle detection and sizing, two optically filtered Xenon flash-lamps are sequentially triggered, providing excitation wavelengths centred at 280 ± 10 and 370 ± 20 nm. The fluorescence emission is collected by two spherical mirrors and split into two channels using a dichroic filter at 410 nm before being measured by two photomultiplier tubes (PMTs).

Both PMTs record fluorescence during the 280 nm excitation phase because no detection bands overlap the excitation band; however only the 410–650 nm PMT detector is active during the 370 nm excitation. In subsequent discussions herein the three fluorescent channels will be referred to as FL1 (fluorescence between 300 and 400 nm, following excitation at 280 nm), FL2 (fluorescence between 410 and 650 nm, following 280 nm excitation) and FL3 (fluorescence between 410 and 650 nm, following excitation at 370 nm). The autofluorescence arising from the 280 nm excitation in biological material is influenced heavily by proteins and the bio-molecule tryptophan, whereas fluorescence from 370 nm excitation is influenced by riboflavin and co-enzyme NAD(P)H (Stanley et al., 2011; Benson et al., 1979; Billinton and Knight, 2001; Foot et al., 2008; Kaye et al., 2005; Li and Humphrey, 1991). However, fluorescence emission spectra are inherently broad, and interrogating complex microorganisms and micron-sized particles results in a complex mixture of fluorescence emission peaks from many fluorophores that can be difficult to interpret unambiguously (Crawford et al., 2014; Pöhlker et al., 2012).

3 Hierarchical cluster methods

Hierarchical agglomerative cluster analysis (HCA) has been demonstrated to be a powerful tool to classify particles (Robinson et al., 2013; Crawford et al., 2014; Gabey et al., 2013); however, the available analysis toolkits are limited by heavy computational burdens, making the analysis of large data sets problematic. In HCA each data point is initially in its own single membered cluster. The clusters are sequentially combined into larger multi-membered clusters until all data points are in one large cluster at the end of the process. At each step through the process the two clusters which are separated by the shortest distance are combined where the inter-cluster distance is determined by the linkage algorithm. In this study we trialled several common linkages, which are now described:

- *Single*: the distance between two clusters is defined as the minimum distance between any single data point in the first cluster and any single point in the second cluster.
- *Complete*: the distance between two clusters is defined as the maximum distance between any single data point in the first cluster and any single point in the second cluster.
- *Average* (unweighted average distance): the distance between two clusters is defined as the average distance between all data points in the first cluster and all data points in the second cluster. The weight of each cluster is proportional to the cluster size.
- *Weighted* (weighted average distance): similar to average, but the weight of each cluster is identical irrespective of size.
- *Ward*: this linkage is a special case where the clusters to be merged is determined by finding the pair of clusters which yield the minimum increase in total within-cluster variance after merging, rather than by minimum distance between clusters.
- *Centroid*: the distance between clusters is defined as the distance between the centres (mean vectors) of clusters.
- *Median*: the distance between two clusters is iteratively defined as the distance between the cluster midpoints. Here the midpoint is defined as the point itself in a singleton cluster or the average of the midpoints of the clusters to be merged.

A full mathematical description of these linkages is provided in Müllner (2013).

3.1 WASP

The WIBS AnalysiS Program (WASP; Robinson et al., 2013) uses the average linkage clustering algorithm and is written in Igor Pro[1]. WASP performs HCA on a random subset of the data limited to a maximum of $\approx 1 \times 10^4$ data points, with analysis taking around 4 h on a high-powered desktop computer[2]. The choice of the number of clusters to retain is manually selected by the inspection of several metrics, and the remaining data are attributed to the chosen clusters by comparison to the cluster centroids using a distance-based similarity method as described in Robinson et al. (2013). The authors noted that this comparative method can lead to systematic misattribution when less populous clusters form poorly defined centroids which do not reflect the true spread of the variables. They also noted that particles outside of a specified distance from a cluster centroid are left unclassified, potentially leading to an underprediction of cluster concentrations.

3.2 Fastcluster

In this manuscript we use open-source HCA methods which can analyse data sets in excess of 1×10^6 points on a desktop computer. Subsampling and comparative attribution are not required as each data point is explicitly clustered. We also test the feasibility of using an automated method for determining the optimum number of clusters to retain.

In this study we have used the open-source Python package fastcluster (Müllner, 2013), which features several common linkages. Of the included linkages the Ward, centroid and median linkages do not require the distances between data points to be stored in memory, allowing for memory-saving modes to be used, greatly increasing the maximum number of data points that can be analysed from approximately 7×10^4 to in excess of 1×10^6 points using the test computer described earlier. In order to take advantage of the memory-saving algorithms, the Euclidean distance metric must be used. The performance of the memory-saving linkages are assessed using laboratory-sampled particles of known type and ambient data previously analysed with WASP. In a future publication we will assess computational requirements in more detail, presenting results pertinent to "big-data" analysis depending on the amount of data retrieved during any given campaign.

3.3 Overview of analysis procedure

In this section we provide an overview of the procedure followed when applying hierarchical agglomerative cluster analysis to WIBS data (summarised in Fig. 1):

1. load and quality assure data;
2. filter data;
   a. remove particles $D_p < 0.8$ µm;
   b. remove non-fluorescent particles;
   c. remove saturating particles.
3. normalise data;
4. cluster data;
5. validate cluster solutions;
6. generate cluster products.

These procedures are now discussed.

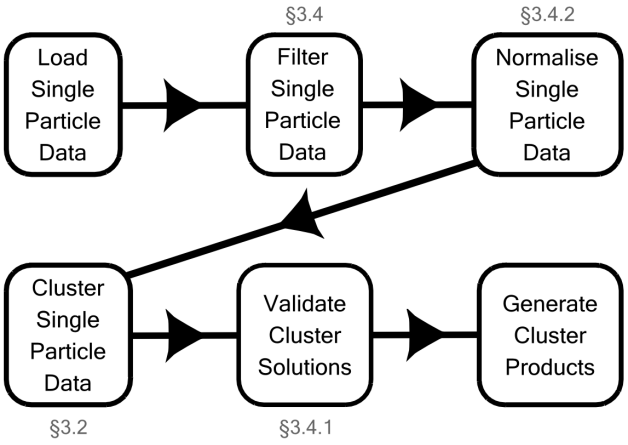

Figure 1. Schematic of procedure followed to generate cluster products from raw data. Relevant sections for each sub-procedure are labelled where appropriate.

3.4 Data preparation

Prior to analysis it is necessary to prepare the single-particle data to ensure that they are physically meaningful to prevent artefacts biasing the cluster solutions such that any potential to effect the performance of any cluster analysis is minimised.

The particle collection efficiency of the WIBS drops below 50 % at ~ 0.8 µm. We have chosen to integrate number concentrations of particles > 0.8 µm rather than apply a correction factor to the concentrations below this size.

The baseline fluorescence of the instrument is measured during so-called forced trigger (FT) sampling periods where the instrument triggers the flash lamps and records the resultant fluorescence in the absence of aerosol in the sample volume. The WIBS-4 instrument automatically makes such measurements if measured concentrations are lower than 2 counts s^{-1} for a sustained period of time, on the basis that the coincidence of a forced trigger measurement with a particle in the measurement region is small. The mean fluorescence in a FT period is treated as the baseline fluorescence of the optical chamber during the sample period. For

[1] WaveMetrics Inc., OR, USA

[2] 3.4 GHz quad core, eight-thread processor, 16 GB RAM, 64 bit OS.

a particle to be considered fluorescent it must exhibit a fluorescence greater than a threshold value, defined as the baseline fluorescence plus 3 SDs (standard deviations). The analysis software subtracts this threshold value from measured fluorescence of each sample, with all values greater than 0 being considered significantly fluorescent compared to the instrument baseline. Fluorescence measurements below the threshold (i.e. less than 0 after threshold subtraction) are not considered physically meaningful and are clipped at 0. This simplifies the segregation between fluorescent (FL) and non-fluorescent (non-FL) particles.

Sufficiently fluorescent particles (such as pollens) will saturate the PMT, and as such it is not possible to accurately measure their true fluorescence. Data from saturating particles are not physically meaningful and are excluded from analysis.

3.4.1 Cluster validation indices

In order to remove the subjective nature of the method employed by Robinson et al. (2013) to determine the optimum number of clusters to retain, we have chosen to use the Calinski–Harabasz criterion (CH), which is defined as being the ratio of the overall between-cluster variance to the overall within-cluster variance (Calinski and Harabasz, 1974). We calculate the CH index for the first 20 cluster solutions and select the solution with the highest CH value as being the optimal solution.

3.4.2 Data normalisation

In the Robinson et al. (2013) study the prepared data were z-score-normalised prior to analysis. This was performed to minimise the effect of the different ranges of scale of each parameter biasing the clustering; i.e. the fluorescent intensities are of the scale 0–2092, size 0.8–20 and AF 0–100. We investigate the effect of normalisation on clustering performance using the following standard procedures:

1. No normalisation.

2. Subtract mean; divide by standard deviation (z-score). The mean value of the normalised distribution is 0, where the z-score value of a data point is the number of standard deviations from the mean. This can be positive or negative.

3. Standardise by range. Subtract minimum value; divide by the maximum value. Normalises data to new range of 0–1.

4. Divide by sum. Divide each of the variables by its sum. The sum of the normalised distribution is 1. Since our original data are positive, the normalised values will also be positive.

Table 1. Properties of the polystyrene latex spheres sampled.

PSL sample	Size [µm]	Doping	Sample size
1	4.17	None	8927
2	3.1	Green	7976
3	2.2	Red	8942
4	2.1	Blue	8796
5	1	Green	5055

5. Rank. Replace each data point by its rank. The data under this normalisation will be integers from 1 to N, where N is the number of data points.

These are the procedures considered in Milligan and Cooper (1988) but excluding procedures which produce identical results for the Euclidean metric. They concluded the range normalisation to be the best performing. We considered procedures proposed by Gnanadesikan et al. (2007) which considered better-performing alternatives to the above procedures. However it seems unlikely that the procedures will scale in terms of performance for large data.

4 Data sets and data preparation

To assess the performance and suitability of the available clustering linkages, we first look at a laboratory data set of known particle types so that the cluster solutions can be compared to the known result. We then trial the best-performing methods on ambient data from the BEACHON-RoMBAS (Bio–hydro–atmosphere interactions of Energy, Aerosols, Carbon, H_2O, Organics and Nitrogen–Rocky Mountain Biogenic Aerosol Study) experiment, which has been studied previously using similar methods (Robinson et al., 2013; Crawford et al., 2014). These data sets are now described in detail.

4.1 Fluorescent polystyrene latex spheres

To test the applicability and performance of the memory-efficient hierarchical agglomerative clustering linkages available in the Python package fastcluster, five different PSL spheres[3] were sampled using the WIBS-4. They were of different sizes, and four of them had been doped with a fluorescent coating. The properties of the tested PSLs are summarised Table 1.

The three fluorescence measurements (FL1–3), size and asymmetry factor were chosen as inputs. The PSLs exhibit strong fluorescence, with some saturating the PMTs in multiple channels; as such we have chosen to keep saturating particles in the analysis to maximise sample size. Non-fluorescent

[3]Manufactured by Polysciences Inc., PA, USA, and Duke Scientific Corp., CA, USA.

Table 2. Performance of the different linkages and normalisation procedures for the full data set in terms of the percentage of data points placed into the same cluster as the known clustering. In bold are the best-performing normalisations for each linkage.

	None	z-score	Range	Sum	Rank
Single	**48.065**	24.384	**48.065**	47.996	42.160
Complete	87.996	**96.039**	87.531	85.126	82.390
Average	87.432	**97.791**	87.406	65.772	96.990
Weighted	85.439	**89.675**	64.843	82.798	65.056
Ward	72.606	**98.136**	98.036	97.726	98.011
Centroid	87.423	**97.264**	87.446	65.772	96.805
Median	82.361	80.575	82.974	**84.912**	65.501

particles and particles smaller than 0.8 µm have been excluded from the analysis due to low collection efficiency. AF and size are typically log-normally distributed. In keeping with the analysis performed in Crawford et al. (2014) and Robinson et al. (2013) we convert these variables to log space prior to analysis. As memory saving is used, this limits analysis to using only the Euclidean distance metric.

4.2 The regional BEACHON-RoMBAS experiment

The WIBS was deployed at the the Manitou Experimental Forest Observatory (MEFO), located 35 km northwest of Colorado Springs, Colorado, USA (Ortega et al., 2014; Kim et al., 2010), as part of the Rocky Mountain Biogenic Aerosol Study project (BEACHON-RoMBAS) during summer 2011. Details of the experiment and sampling arrangement are given in Crawford et al. (2014). In the Crawford et al. (2014) study HCA was performed on a subset of the WIBS data ($\approx 1 \times 10^4$ particles) using the average linkage, with the remaining particles attributed to a cluster by comparison to the cluster centroid. Details of the attribution method and the process of selecting the number of clusters to retain are provided in Robinson et al. (2013). This analysis yielded clusters which were behaviourally consistent with fungal spores and bacteria. We perform analysis of this data set using the methods described in this manuscript, which we compare to the Crawford et al. (2014) results.

5 Results

5.1 Fluorescent polystyrene latex spheres

The fastcluster package was run with the seven available linkages, each with the different normalisation procedures. Note that only the single, Ward, centroid and median linkage are available when the memory-efficient version of the fastcluster package is used.

Table 3. Performance of the Ward linkage for varying sample size.

Sample size	500	1000	5000	10 000	20 000
z-score	79.330	85.696	94.746	97.543	97.132
Range	95.664	97.671	98.041	98.065	98.074

We use the CH index to identify the "optimal" number of clusters and attempt to construct a best match between the desired clusters and proposed clusters. Then, to evaluate the performance of the algorithm, we calculate the proportion of the data points placed into the same cluster for both the desired and proposed clustering. The results are given in Table 2.

For the full data set we can see that the z-score is the best-performing normalisation for all but the single and median linkages, where the performance is poorer across all normalisations.

However in Table 3 we repeat the tests for varying sample size, where we see that as sample size decreases the range normalisation starts to outperform the z-score.

It appears that when using the full data set the z-score normalisation with either the Ward linkage or average linkage is the preferred option. When sampling, however, we see that range normalisation may be better.

An explanation for this behaviour could be that the range normalisation suffers with outliers which we are much more likely to encounter for large samples, so we would expect better performance for the smaller samples. Contrast this with the z-score where our measurement of the mean and the standard deviation is more accurate with large samples.

Figure 2 shows the cluster centroids for the Ward linkage with range and z-score normalisation. It can be seen that both methods yield similar clusters to the known solution; e.g. the average values of the 4.17 µm sample are accurately captured by the fifth cluster using range normalisation and the third cluster using z-score normalisation. Similarly the 3.1 µm green PSL sample is captured by the fourth range-normalised and first z-score-normalised clusters. Figures 3 and 4 show a time series of the FL1–3 and size input parameters (AF omitted from figure), which are colour-coded by the cluster assignments in Fig. 2. The bottom panel of each figure shows the fraction of each cluster assigned to each sample, where it can be seen that both normalisation methods achieve a high level of attribution accuracy, with a minimum of 96 % of data points being correctly attributed with no significant misattribution. The results of this experiment suggest that both range and z-score normalisation are appropriate when clustering WIBS data using the Ward linkage, with each yielding an optimal five-cluster solution correctly attributing 98 and 98.1 % of the data points respectively. The centroid linkage with z-score normalisation also

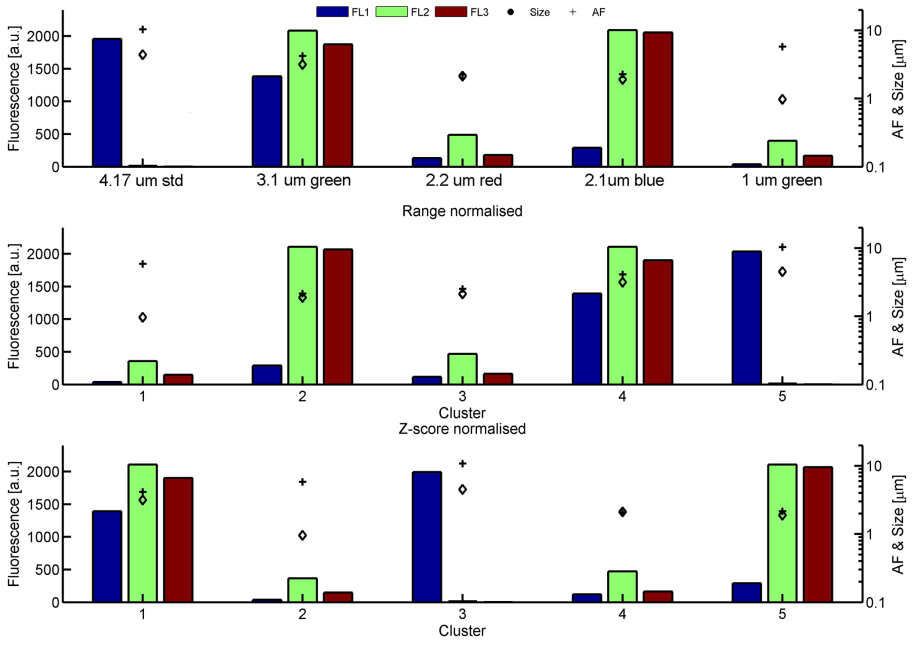

Figure 2. Top panel: average FL1–3 detector intensites (blue, green and brown bars, left axis), size (diamond, right axis) and asymmetry factor (cross, right axis) for the five PSL samples. Middle and bottom panels: the same as for the top panel but for the Ward linkage solution centroids using range (middle) and z-score (bottom) normalisation.

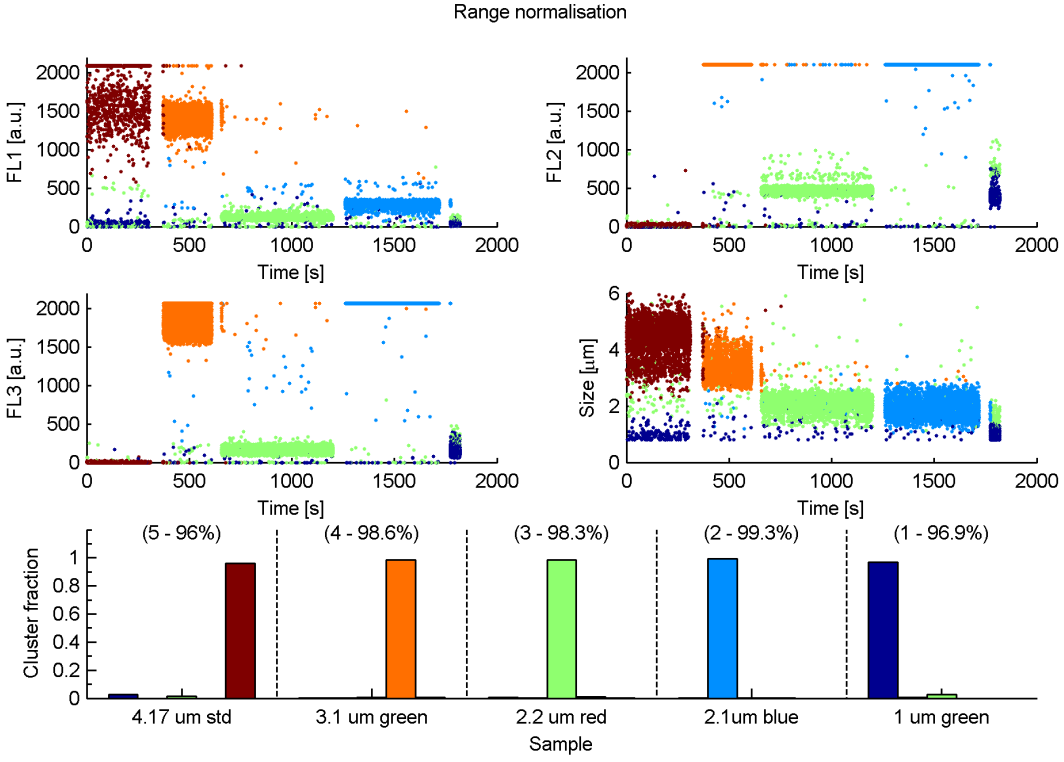

Figure 3. Time series of PSL samples with data points coloured by cluster assignment for Ward linkage and range normalisation. Bottom panel shows the fraction of each cluster assigned to each sample with the most populated cluster annotated above.

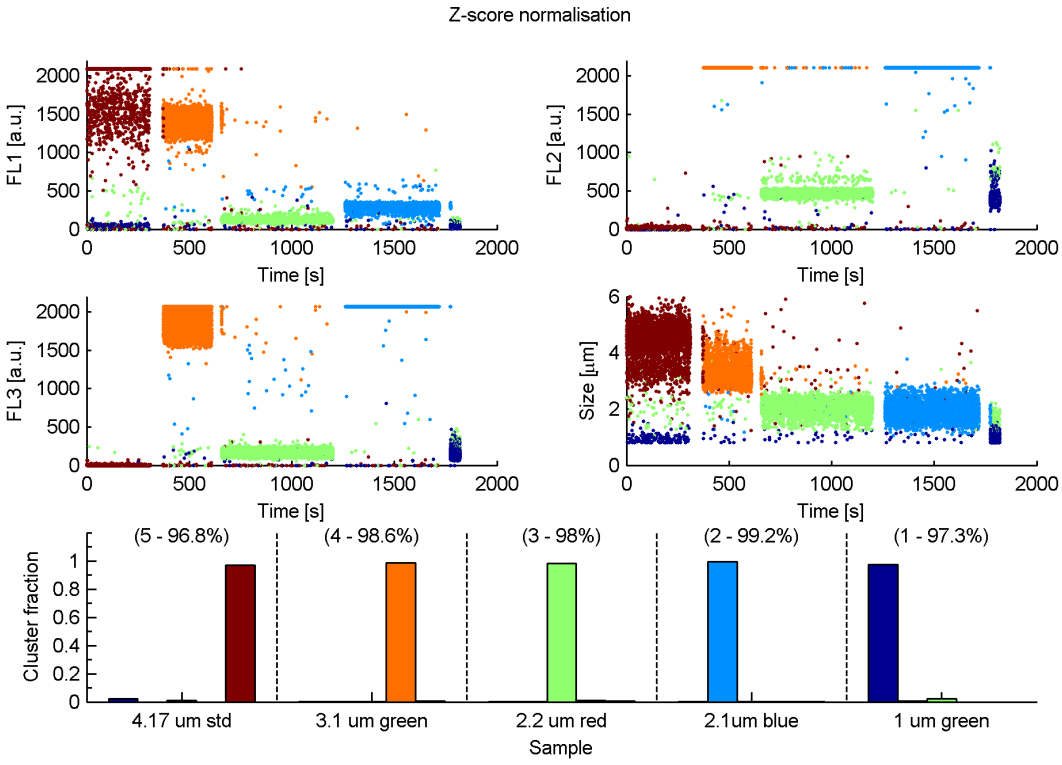

Figure 4. Same as Fig. 3 but for Ward linkage and z-score normalisation.

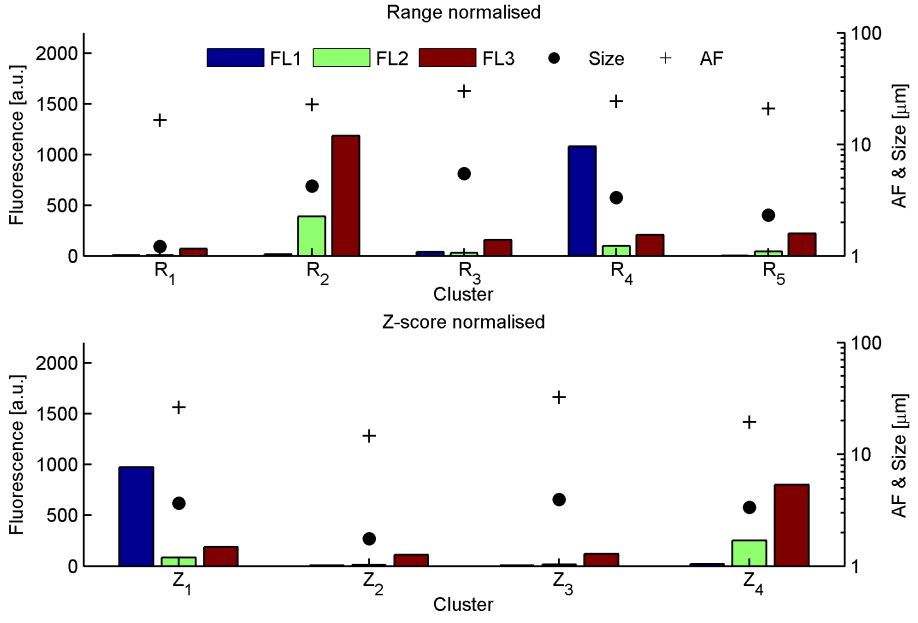

Figure 5. Same as Fig. 2 but for BEACHON-RoMBAS ambient data.

performed well, correctly attributing 97.3 % of the particles into five significant clusters.

5.2 BEACHON-RoMBAS

Data from the BEACHON-RoMBAS experiment ($\approx 8.2 \times 10^5$ fluorescent data points) were analysed using the Ward linkage with both range and z-score nor-

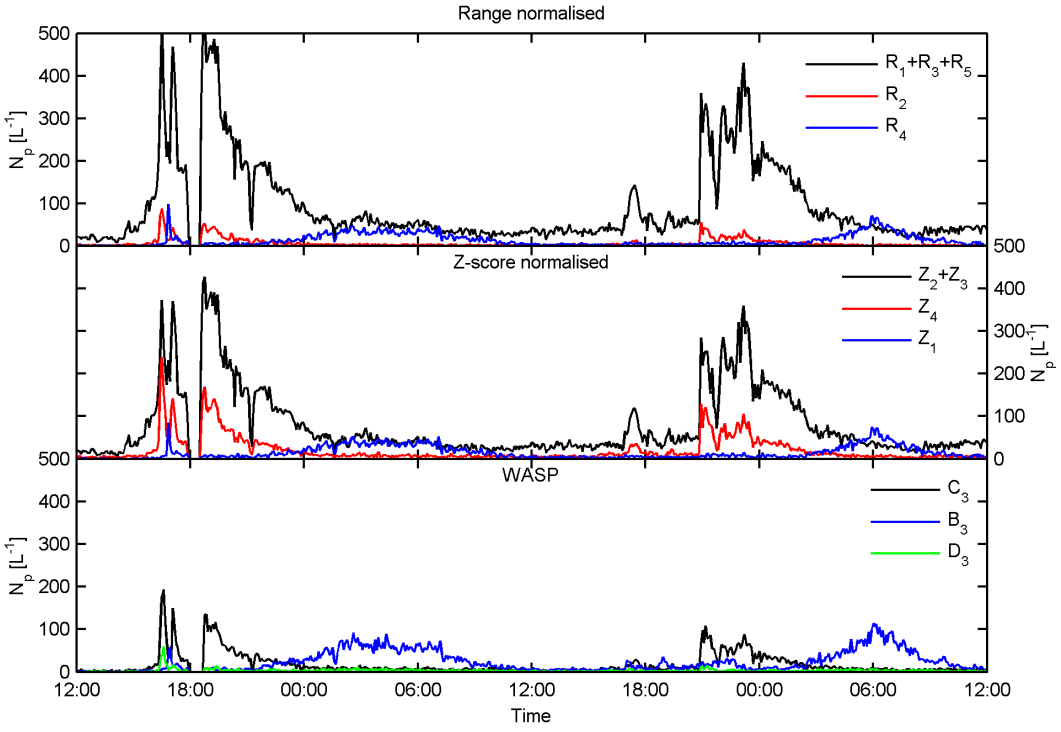

Figure 6. Time series of BEACHON-RoMBAS cluster concentrations using Ward linkage with range (top panel) and z-score (middle panel) normalisation as compared to the solutions obtained using WASP (bottom panel) for the period 00:00 MST (Mountain Standard Time) on 26 July 2011 to 12:00 MST on 28 July 2011. Clusters with similar centroids have been combined. See text for details.

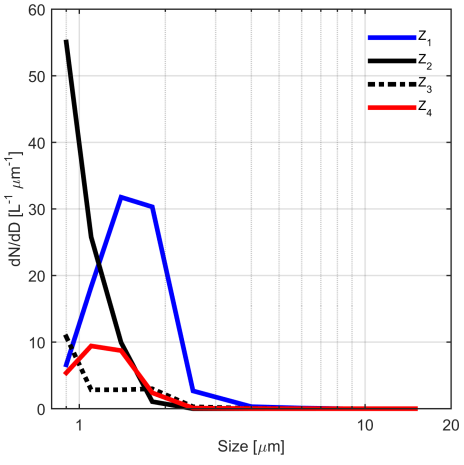

Figure 7. Size distribution of BEACHON z-score-normalised clusters produced using the Ward linkage for the period 00:00 to 06:00 MST 27 July 2011.

malisations and also the centroid linkage with z-score normalisation. The centroid linkage yielded a solution with only one significantly populated cluster, suggesting that it is inappropriate for analysing ambient data. Figure 5 shows the cluster centroids of each Ward normalisation where the range yields a five-cluster solution and z-score a 4-cluster solution. It can be seen that the solutions of each are broadly similar; range cluster 4 (hereby notated as R_4) is similar to z-score cluster 1 (hereby notated as Z_1); R_2 is similar to Z_4. Additionally R_1, R_3 and R_5 are similar to R_2, suggesting that they are of similar origin, with the difference in fluorescence being due to size, morphology or particle age. This is also observed in the z-score result in clusters Z_2, Z_3 and Z_4. A time series (not shown) of cluster concentrations shows these internally similar clusters to respond in a similar fashion to meteorological events such as rainfall. For ease of interpretation the concentrations of similar clusters have been combined. Figure 6 shows a time series of the combined cluster concentrations for each method and also the cluster concentrations obtained using WASP. It can be seen that the concentrations of clusters $R_1 + R_3 + R_5$, R_2, $Z_2 + Z_3$ and Z_4 all behave in a similar fashion to the WASP cluster C_3, which was determined to be representative of bacteria owing to its strong positive response to rainfall (Crawford et al., 2014). The response of clusters R_4 and Z_1 is similar to the WASP cluster B_3, which was determined to be representative of fungal spores owing to its diurnal response to relative humidity. The size distributions for each of the z-score clusters (Fig. 7) show the bacterial clusters to be small with sub-micron modes for clusters Z_2 and Z_3, which is consistent with the bacteria observed at the site, while the fungal cluster (Z_1) mode is approximately 1.5–2 µm as might be expected. Caution must be taken when

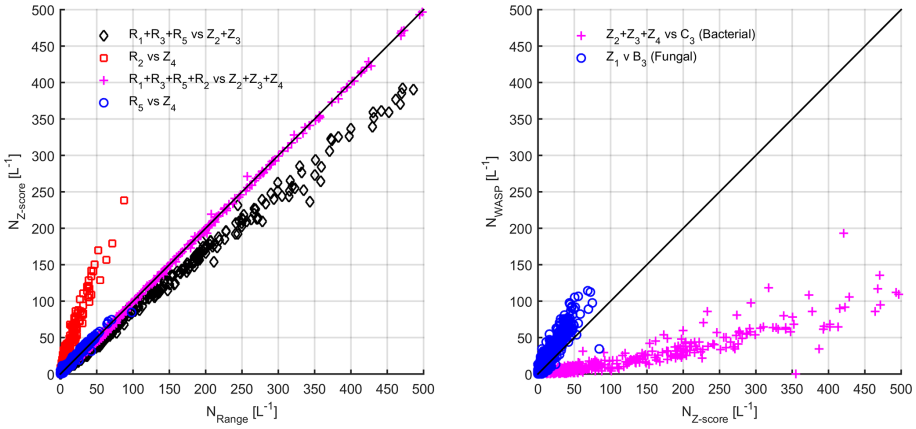

Figure 8. Left panel: comparison of Ward linkage cluster concentrations using range and z-score normalisation for BEACHON-RoMBAS. Right panel: comparison of Ward linkage cluster concentrations (z-score normalisation) to WASP cluster concentrations.

interpreting or assigning a bioaerosol meta-class to a cluster to avoid conflation of different particle types; e.g. emissions of some fungal spore species are positively correlated rainfall, which could be conflated into the bacterial meta-class in this case. Supporting measurements are needed to determine which species are present so that possible conflations can be identified and caveated appropriately.

Figure 8 compares the concentrations of the similar clusters for each normalisation method. Comparison of R_5 to Z_4 (left panel, blue circles, representative of fungal spores) shows each method to yield similar concentrations. Comparison of the bacterial cluster concentrations yields poor correlation between methods when comparing the traces in Fig. 6 (left panel, black diamonds and red squares); however when the concentration of all clusters representative of bacteria are combined (left panel, magenta crosses) the correlation is excellent ($N_{z\text{score}} = 1.00 \times N_{\text{range}} - 1.42$, $R^2 = 1$). This suggests that the major difference between the two different normalisation methods is how particles of similar types are partitioned between the clusters.

The right panel of Fig. 8 compares the z-score concentrations to those obtained using WASP. It can be seen that the WASP fungal concentration is overestimated by a factor of approximately 1.5 compared to the z-score result (Z_4 and B_3, blue circles). The WASP bacterial concentration is underestimated by approximately a factor of 5 compared to the z-score result. Figure 9 shows the hourly average diurnal cycles of the fungal (top panels) and bacterial (bottom panels) cluster concentrations for the z-score result (left panels) and WASP (right panels) over the period 27 July 2011–7 August 2011. Each method displays a similar trend, with the fungal clusters exhibiting a minimum during the day owing to the diurnal response of fungal spores to relative humidity and the bacterial clusters responding to the frequent afternoon rain storms (Crawford et al., 2014). Again it can be seen that WASP overestimates the fungal concentration by approximately a factor of 1.5–2 and underestimates the bacterial concentration by a factor of 5–6 compared to the z-score result. The most likely explanation for the observed discrepancies between the WASP and z-score concentrations is the introduction of artefacts caused by the subsampling and comparative attribution methods used in WASP. In the fungal spore case, misattribution due to a poorly defined centroid can lead to an overestimation when compared to the new method as observed here. WASP yields only one cluster representative of bacteria, while the z-score method yields three and the range method four. This results in WASP failing to attribute data points potentially representative of bacteria to its single bacterial cluster, leading to the observed underestimation when compared to the new method. WASP does not return diagnostic information about the cluster attribution; however, the sum of the concentration of WASP clusters B_3, C_3 and D_3 only accounts for approximately 24 % of the fluorescent aerosol concentration, suggesting that many particles are left unattributed by WASP.

6 Conclusions

Several hierarchical agglomerative cluster analysis linkages and normalisation methods were trialled using several laboratory samples of known particle type and a previously published ambient data set which was analysed using similar methods. The Ward linkage with range and z-score normalisation was found to successfully resolve the five test PSL samples with a high level of accuracy, correctly attributing 98 and 98.1 % of the data points respectively. Analysis of the BEACHON-RoMBAS WIBS-3 data yielded similar results using the Ward linkage with the range and z-score normalisation methods. Each method produced one cluster representative of fungal spores and several clusters representative of bacterial aerosol where the fungal concentrations and the sum of the bacterial aerosol concentrations agreed well. The BEACHON-RoMBAS results were compared to the WASP

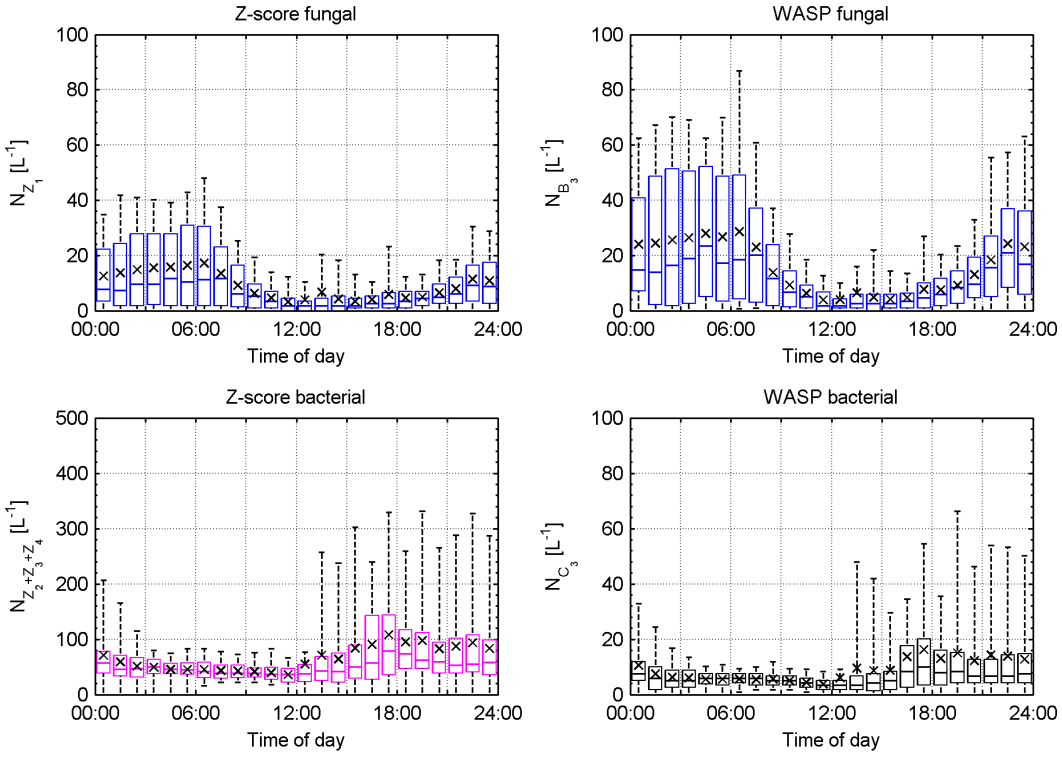

Figure 9. Top panels: hourly average diurnal cycle of fungal cluster concentration for z-score normalisation (left panels) and WASP (right panels) over the period 27 July 2011–7 August 2011. Bottom panels: same as for top panels but for the bacterial clusters. Whiskers denote 5th and 95th percentiles. Mean value indicated by x marker. Note change in scale for bacterial panels.

results for the same data set (Robinson et al., 2013; Crawford et al., 2014), where it was found that WASP overestimated the fungal spore concentration by a factor of 1.5 and underestimated the bacterial aerosol concentration by a factor of 5 compared to the methods trialled here. This is likely due to errors arising from misattribution due to poor centroid definition and failure to assign particles to a cluster as a result of the subsampling and comparative attribution method employed by WASP. The methods used here allow for the entire fluorescent population of particles to be analysed, yielding an explicit cluster attribution for each particle. This improves cluster centroid definition (e.g. allowing for several bacterial clusters compared to just one in WASP) and removes the potential for underestimation by failing to attribute a particle to a cluster.

In this paper we have demonstrated that WIBS single-particle UV-LIF spectrometer data can be successfully segregated using the Ward hierarchical agglomerative cluster analysis linkage with z-score and range data normalisation. The explicit clustering method employed in this study can be applied to large data sets, removing potential clustering aretfacts associated with the subsampling and attribution method used in previous approaches, improving our capacity to discriminate and quantify PBAP meta-classes. These improved techniques will be of importance for interpreting data from future multi-parameter UV-LIF instruments with improved fluorescence resolution and for extending the measurement technique to real-time quantification for ambient monitoring networks.

The Supplement related to this article is available online at doi:10.5194/amt-8-4979-2015-supplement.

Acknowledgements. The authors wish to express our gratitude to R. Sarda Esteve (CEA) and J. A. Huffman (University of Denver) for use of their fluorescent calibration particles as part of the BIODETECT experiment, without which the fundamentals of this work could not have been performed. We would also like to acknowledge USFS and NCAR for providing invaluable logistical support and access to the Manitou Experimental Forest field site. We also acknowledge P. Kaye and W. R. Stanley (University of Hertfordshire) for their continued support. This work was funded by the Natural Environment Research Council INUPIAQ programme, grant number NE/K006002/1.

Edited by: F. Pope

References

Benson, R., Meyer, R., Zaruba, M., and KcKhann, G.: NoCellular autofluorescence – is it due to flavins?, J. Histochem. Cytochem., 27, 44–48, 1979.

Billinton, N. and Knight, A. W.: Seeing the wood through the trees: a review of techniques for distinguishing green fluorescent protein from endogenous autofluorescence, Anal. Biochem., 291, 175–97, doi:10.1006/abio.2000.5006, 2001.

Calinski, T. and Harabasz, J.: A dendrite method for cluster analysis, Commun. Stat.-Theor. M., 3, 1–27, doi:10.1080/03610927408827101, 1974.

Crawford, I., Bower, K. N., Choularton, T. W., Dearden, C., Crosier, J., Westbrook, C., Capes, G., Coe, H., Connolly, P. J., Dorsey, J. R., Gallagher, M. W., Williams, P., Trembath, J., Cui, Z., and Blyth, A.: Ice formation and development in aged, wintertime cumulus over the UK: observations and modelling, Atmos. Chem. Phys., 12, 4963–4985, doi:10.5194/acp-12-4963-2012, 2012.

Crawford, I., Robinson, N. H., Flynn, M. J., Foot, V. E., Gallagher, M. W., Huffman, J. A., Stanley, W. R., and Kaye, P. H.: Characterisation of bioaerosol emissions from a Colorado pine forest: results from the BEACHON-RoMBAS experiment, Atmos. Chem. Phys., 14, 8559–8578, doi:10.5194/acp-14-8559-2014, 2014.

Douwes, J., Thorne, P., Pearce, N., and Heederik, D.: Bioaerosol Health Effects and Exposure Assessment: Progress and Prospects, Ann. Occup. Hyg., 47, 187–200, doi:10.1093/annhyg/meg032, 2003.

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: Optically Based Biological and Chemical Detection for Defence, 71160I–71160I-12, 15 September 2008, Cardiff, Wales, UK, doi:10.1117/12.800226, 2008.

Gabey, A. M.: Laboratory and field characterisation of fluorescent and primary biological aerosol particles, PhD thesis, University of Manchester, Manchester, UK, 2011.

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.

Gabey, A. M., Vaitilingom, M., Freney, E., Boulon, J., Sellegri, K., Gallagher, M. W., Crawford, I. P., Robinson, N. H., Stanley, W. R., and Kaye, P. H.: Observations of fluorescent and biological aerosol at a high-altitude site in central France, Atmos. Chem. Phys., 13, 7415–7428, doi:10.5194/acp-13-7415-2013, 2013.

Gnanadesikan, R., Kettenring, J., and Maloor, S.: Better alternatives to current methods of scaling and weighting data for cluster analysis, J. Stat. Plan. Infer., 137, 3483–3496, doi:10.1016/j.jspi.2007.03.026, 2007.

Heald, C. L. and Spracklen, D. V.: Atmospheric budget of primary biological aerosol particles from fungal spores, Geophys. Res. Lett., 36, L09806, doi:10.1029/2009GL037493, 2009.

Hummel, M., Hoose, C., Gallagher, M., Healy, D. A., Huffman, J. A., O'Connor, D., Pöschl, U., Pöhlker, C., Robinson, N. H., Schnaiter, M., Sodeau, J. R., Stengel, M., Toprak, E., and Vogel, H.: Regional-scale simulations of fungal spore aerosols using an emission parameterization adapted to local measurements of fluorescent biological aerosol particles, Atmos. Chem. Phys., 15, 6127–6146, doi:10.5194/acp-15-6127-2015, 2015.

Jacobson, M. Z. and Streets, D. G.: Influence of future anthropogenic emissions on climate, natural emissions, and air quality, J. Geophys. Res., 114, D08118, doi:10.1029/2008JD011476, 2009.

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.

Kaye, P. H., Aptowicz, K., Chang, R. K., Foot, V., and Videen, G.: Angularly Resolved Elastic Scattering from Airborne Particles, Opt. Biol. Part., 238, 31–61, 2007.

Kim, S., Karl, T., Guenther, A., Tyndall, G., Orlando, J., Harley, P., Rasmussen, R., and Apel, E.: Emissions and ambient distributions of Biogenic Volatile Organic Compounds (BVOC) in a ponderosa pine ecosystem: interpretation of PTR-MS mass spectra, Atmos. Chem. Phys., 10, 1759–1771, doi:10.5194/acp-10-1759-2010, 2010.

Li, J. K. and Humphrey, A. E.: Use of fluorometry for monitoring and control of a bioreactor, Biotechnol. Bioeng., 37, 1043–1049 doi:10.1002/bit.260371109, 1991.

Milligan, G. W. and Cooper, M. C.: A study of standardization of variables in cluster analysis, J. Classif., 5, 181–204, doi:10.1007/BF01897163, 1988.

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.

Morris, C. E., Conen, F., Alex Huffman, J., Phillips, V., Pöschl, U., and Sands, D. C.: Bioprecipitation: a feedback cycle linking earth history, ecosystem dynamics and land use through biological ice nucleators in the atmosphere, Global Change Biol., 20, 341–351, doi:10.1111/gcb.12447, 2014.

Müllner, D.: fastcluster: fast hierarchical, agglomerative clustering routines for R and Python, J. Stat. Softw., 9, 1–18, doi:10.18637/jss.v053.i09, 2013.

Ortega, J., Turnipseed, A., Guenther, A. B., Karl, T. G., Day, D. A., Gochis, D., Huffman, J. A., Prenni, A. J., Levin, E. J. T., Kreidenweis, S. M., DeMott, P. J., Tobo, Y., Patton, E. G., Hodzic, A., Cui, Y. Y., Harley, P. C., Hornbrook, R. S., Apel, E. C., Monson, R. K., Eller, A. S. D., Greenberg, J. P., Barth, M. C., Campuzano-Jost, P., Palm, B. B., Jimenez, J. L., Aiken, A. C., Dubey, M. K., Geron, C., Offenberg, J., Ryan, M. G., Fornwalt, P. J., Pryor, S. C., Keutsch, F. N., DiGangi, J. P., Chan, A. W. H., Goldstein, A. H., Wolfe, G. M., Kim, S., Kaser, L., Schnitzhofer, R., Hansel, A., Cantrell, C. A., Mauldin, R. L., and Smith, J. N.: Overview of the Manitou Experimental Forest Observatory: site description and selected science results from 2008 to 2013, Atmos. Chem. Phys., 14, 6345–6367, doi:10.5194/acp-14-6345-2014, 2014.

Pöhlker, C., Wiedemann, K. T., Sinha, B., Shiraiwa, M., Gunthe, S. S., Smith, M., Su, H., Artaxo, P., Chen, Q., Cheng, Y., Elbert, W., Gilles, M. K., Kilcoyne, A. L. D., Moffet, R. C., Weigand, M., Martin, S. T., Pöschl, U., and Andreae, M. O.: Biogenic potassium salt particles as seeds for secondary organic aerosol in the Amazon, Science, 337, 1075–1078, doi:10.1126/science.1223264, 2012.

Robinson, N. H., Allan, J. D., Huffman, J. A., Kaye, P. H., Foot, V. E., and Gallagher, M.: Cluster analysis of WIBS single-particle bioaerosol data, Atmos. Meas. Tech., 6, 337–347, doi:10.5194/amt-6-337-2013, 2013.

Sands, D., Langhans, V., Scharen, A., and de Smet,: The association between bacteria and rain and possible resultant meteorological implications, J. Hungar. Meteorol. Serv., 86, 148–152, 1982.

Schumacher, C. J., Pöhlker, C., Aalto, P., Hiltunen, V., Petäjä, T., Kulmala, M., Pöschl, U., and Huffman, J. A.: Seasonal cycles of fluorescent biological aerosol particles in boreal and semi-arid forests of Finland and Colorado, Atmos. Chem. Phys., 13, 11987–12001, doi:10.5194/acp-13-11987-2013, 2013.

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.

Toprak, E. and Schnaiter, M.: Fluorescent biological aerosol particles measured with the Waveband Integrated Bioaerosol Sensor WIBS-4: laboratory tests combined with a one year field study, Atmos. Chem. Phys., 13, 225–243, doi:10.5194/acp-13-225-2013, 2013.