We would like to thank the reviewers for their comments. We apologize for the extended time between reviews and revision but this was due to incorporating the important extra work requested in the revised manuscript. We have made the changes suggested and responded to comments in a point by point format below. We believe this is a stronger paper as a result and thank the reviewers for their work.

Reviewer #1

1.) The main scientific work and ideas that were put into the paper are the growing of the forests and their validation including variable reduction and creation of the confusion matrix. Accordingly, four out of five figures are about these topics. And the results are interesting and offer new ways of looking at this kind of datasets. In contrast the abstract, mainly the introduction and the conclusion strongly focus on the prediction of aerosol classes.

In response to this reviewer comment we have added text in the three stated areas (abstract, intro and conclusions) to the emphasize the use of the general technique as opposed to the specific classification of aerosol spectra. For example, we now lead the introduction with the generalized technique and then move to the specific use here. We believe it is also critical to maintain the classification objective as this was our motivation for the work and the data we present.

> 2: 8-12 in the abstract now reads:
> *"Our primary focus surrounds the growing of random forests using feature selection to reduce dimensionality, and the evaluation of trained models with confusion matrices. In addition to classifying 20 unique but chemically-similar aerosol types, models were also created to differentiate aerosol within four broader categories: fertile soils, mineral/metallic particles, biological, and all other aerosols."*
>
> 3: 1-15 in the introduction:
> *"Following the introduction of random forests in the 1990s, recent developments in deep learning and neural networks have triggered a renewed interest in machine learning. This has led to the development of numerous easy-to-use, freely-available, open-source packages in popular programming languages like Python, and these tools are becoming increasing used in academia and industry. While random forests have been used for complex classification and regression analysis in various fields, studies that employ random forests in aerosol mass spectrometry remain sparse. Utilizing these tools, the primary purpose of our study is to introduce a framework for growing random forests, reducing dimensionality, ranking chemical features, and evaluating performance using confusion matrices. Such properties are desirable for SPMS studies, where input variables can become redundant and interpretability is more limited methods with methods such as cluster analysis and neural networks. Powerful analysis techniques such as those falling out of recent artificial intelligence research can prove useful for*

helping to tease out the subtle yet significant impact that aerosol chemistry has on the climate system."

19: 16-21 in the conclusion:

"This study lays out a framework for training and implementing random forests on SPMS data, with a focus on dimensionality reduction and the evaluation of model performance with confusion matrices. A key benefit to the proposed method is chemical feature selection, which allows researchers to identify potentially important chemical markers between arbitrary groups of aerosols or identify sources of contamination. Additionally, the approach allows for differentiation of aerosols within a SPMS dataset, augmenting existing tools and reducing the need for a qualitative comparison between mass spectra."

2.) The paper is in its present form hard to follow. Often the nomenclature is not consistent throughout the paper or doesn't fit to the cited literature. For example, they do not use the proper term "random forest" but call it machine learning classifier, predictive model, classification model, rule-based probabilistic classification of a decision tree ensemble or supervised classification and even more important.

We thank the reviewer for the suggestion and agree more specific language should be applied to describe the random forest algorithm, which is the primary focus of the paper. The terminology through the paper has been modified to be consistent and precise per the comments, and we have removed sections that went into unnecessary detail.
While do note the terms mentioned are appropriate for introducing and motivating the random forest approach, they describe broader categories of machine learning models. The distinction between each was not necessary for this paper, beyond the mention of how supervised methods (i.e. classification via random forests) differ from unsupervised methods such as clustering. We have now tried to make this distinction and believe it is now more clear.

3.) While the basic algorithm to grow a random forest is presented. The underlying concepts (randomness, law of big numbers, assumptions, input parameter) and details of the validation process remain unclear…

As suggested, we have modified the text in several ways to address this comment, which has been broken into specific points and addressed:

Upon further inspection, the authors agree the concepts you mention such as randomness and out-of-bag sampling are key details surrounding random forests approach and needed further detail. Out-of-bag samples refers to spectra that are held out during the validation process to prevent training and testing on the same data, and the section has been revised to more concisely explain the procedure.

12: 2-10 now reads:

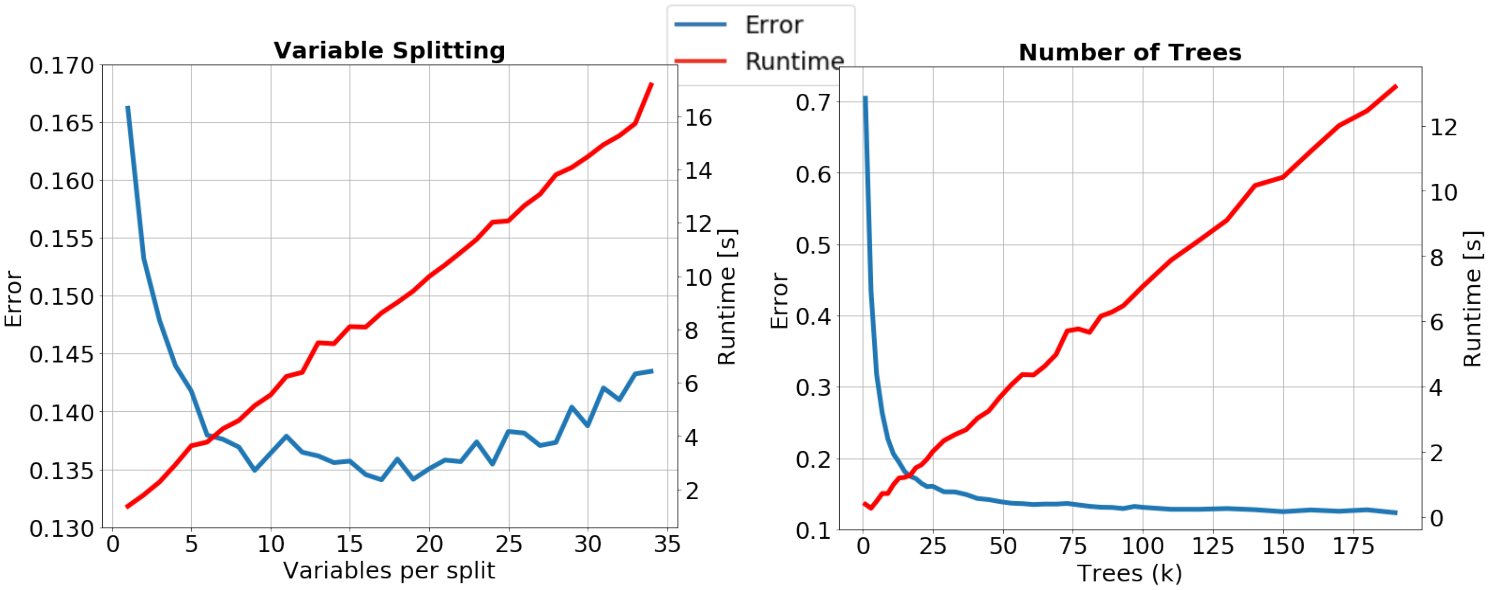
"Overall, the generalizability and robust performance of random forests is owed significantly to the series of random statistical procedures used to construct such models. An ensemble classifier reduces variability by averaging predictions over a series of independently trained models, and bagging introduces additional randomness by producing "perturbed" versions of the original data via random sampling of input data. The randomness used in constructing forests, both in bagging the training set and choosing variable splits, work to decorrelate the output of each tree even as the inputs become correlated [Breiman, 2001]. As the number of trees increases, the law of large numbers guarantees a convergence of the out-of-bag error to the generalization error."

…For example, k=1000 trees have been used for each forest but no further explanation is given why exactly this number of trees is the right one. Or a plot of the test set error against number of trees presented which would make this decision obvious. The number of random variables used to select the best split from is not specified nor its implications discussed….

- We have included various pieces of the information you requested in the paper, and agree the details surrounding runtime, memory requirements, and model selection are important. Additionally, supplementary plots that characterize how model performance and runtime scale are provided below. The plot also shows how runtime scales with number of trees and number of split variables.

11:12 - 20 now includes the requested information:

"The number of variables per split is chosen to be 11 and the number of trees is 110. The optimal model was determined by enumerating combinations of these parameters on a coarse grid and selecting the values that produce the lowest test error. Model behavior is primarily sensitive to the number of variables per split, and shows weak dependence on the number of trees and number of input variables beyond small values. As the number of variable splits increases, error decreases exponentially to a local minimum before again rising due to over fitting. Alternatively, as the number of trees is increased the error asymptotes to some nonzero value, a known characteristic of random forests where test error converges to the generalization error.



…The treatment of the "out-of-bag" observations, which is the central means of validation is not comprehensible….

More detail has been added.

10:15 - 10:22 now reads:

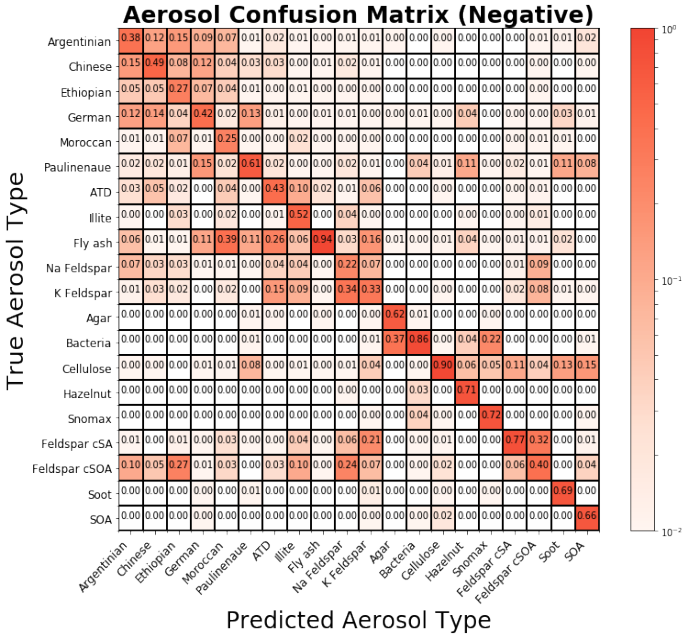
"On average, each tree is built with ~63% of the original data, leaving a portion of the training set unsampled. The unsampled data for each tree, known as 'out-of-bag' observations, are recorded and later provide a means to assess classification error for the forest. To determine model error, predictions are made on each point in the dataset using only the subset of trees that did not use the point for training. Each training point is left out at least once. This is analogous to making predictions with a separately trained forest that did not observe the point and prevents testing with the same data used for training."

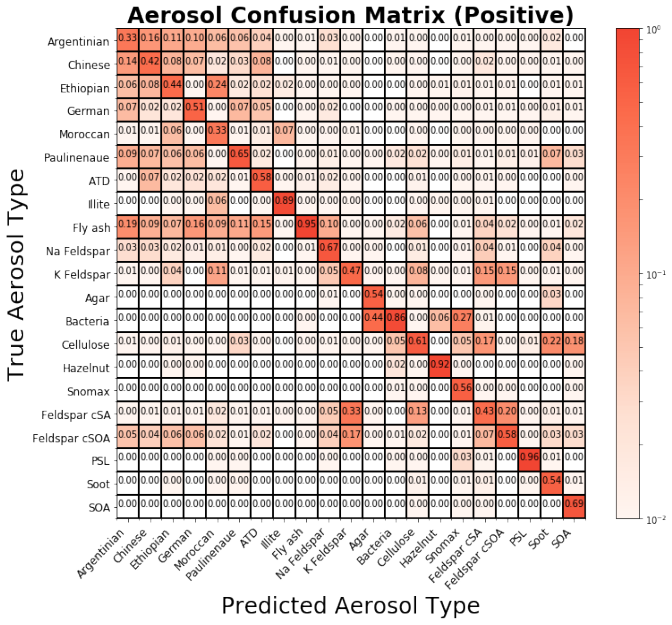
…The resultant classification accuracies are not put into perspective; thus the reader can't judge if the algorithm performs is a major improvement to other methods, of which the simplest would be to just use mean values of each aerosol class and use the most similar one as a prediction…

We have compared the technique to a simple classifier that uses the euclidean distance to assigned an unknown aerosol to the closest "mean" class vector. Confusion matrices for the broad categories have been included in the paper for comparison as part of figure 4, and matrices for all labels have been included below as a supplementary figure.

13:10 - 20 now reads:

"To access relative model performance, we contrast the results with a simple classifier that compares unseen aerosols to a set of class mean vectors. Using the Euclidean distance metric, the unknown aerosol is assigned to the nearest class. This simple baseline classifier helps put results in the context of machine learning techniques that rely on distance-based metrics such as k-means and hierarchical clustering. K-means clustering attempts to divide the data points into k distinct clusters, representing spectra as vectors. Using Euclidean distance, the standard algorithm assigns points to centroids, or clusters, which are essentially mean vectors representing the average of all points in the cluster. Assuming perfect convergence of k-means clustering, where k is the number of aerosol classes, each cluster represents the mean of aerosol in that class. The random forest results below demonstrate many areas of improvement over the simple classifier."





…It is not given which implementation of the algorithm is used. Nor how long a typical random forest generation lasts and how this runtime scales with respect to number of particles, number of trees, number of split variables, etc. . Along with the memory requirements which are missing too, these are basic and easy to provide information that help to compare this method to other methods….

11:20 – 12:1

"The models were trained with the Python 2.7 Scikit-learn module on a MacBook Pro with 16 GB 1600 MHz DDR3 memory and a 2.5 GHz Intel Core i7 processor. A typical

random forest model took about 5-10 seconds to train, and we found a linear relationship between runtime and both the number of trees and variables per split."

4.) The random forests have been grown on chemical information and the size of individual aerosol particles, but some of the aerosol classes are not chemically defined. (e.g. multiple fertile soil classes, ATD) This basic contradiction is not clearly addressed.

We agree with the reviewer and have expanded the paper to define this more clearly. This is in the form of the new paragraph in the introduction on page 5 which also includes detail that some of this issue stems from the complex nature of atmospheric aerosols that are often defined by source as opposed to type. Furthermore, we have color coded Table 1 to make the distinction more clear.

4:16 – 5:9

"Chemical composition of an individual atmospheric aerosol particle is a complex interplay between its primary composition at the source (i.e. dust, biogenic organic, anthropogenic organic, soot, etc.) and its atmospheric processing up to the time of detection. Atmospheric processing can include any combination of coating with secondary material, coagulation and cloud processing. Even distinct primary aerosol types can have similar mass spectral markers. For example, fly ash, mineral dust and bioaerosol can all contain strong phosphate signal [Zawadowicz et al., 2017]. Secondary material is often difficult to differentiate from primary material, but even minor compositional changes can be atmospherically important. As one example, mineral dusts are known to be effective at nucleating ice clouds [Cziczo et al., 2013]; however, despite minor addition of mass, atmospherically processed mineral dust is less suitable for ice formation [Cziczo et al., 2013]. As a second example, ice nucleation in mixed-phase clouds has been suggested to be predominantly influenced by feldspar, a single component among the diverse mineralogy of atmospheric dust [Atkinson et al., 2013]. Using current SPMS data analysis approaches, it can be difficult to detect these minor yet important compositional differences and new robust and generalizable analysis techniques are critical."

5.) To me the section dealing with the blind test data does not fit to the abstract and introduction which present the random forest as a tool specifically suited for this use-case. After showing 80+

There are two primary factors that help explain differences between the test set and blind set, which are both due to the way the experiment and sampling were conducted : a) transmission efficiency b) coagulation. During the course of the experiment, we expect the mineral dust and SOA to coagulate. Since aerosol types were reported by AIDA before particles enter the chamber, it is not possible to quantify exactly what fraction of the particles picked up an SOA coating. Moreover, there would have been a time dependence to the coagulation process.

Additionally, through coagulation, there is the possibility of effectively producing a particle type not in the training set, depending on the exact mineral component of the mineral dust used by AIDA. While it is known a mineral dust was included in the chamber, the exact composition of the dust was not known. While our training set contains K-Feldspar coated with SOA, a different type of SOA-coated mineral dust will appear unique to the model. Because the generalization performance of supervised classifiers is ill-defined for particles not included in the training set, this could lead to performance that is not captured by the confusion matrices. Given the experimental uncertainties from transmission efficiency and coagulation, as well as the model uncertainties highlighted in the confusion matrices, we believe the results reveal skill in using random forests to pick out distinct aerosols. In future studies, uncertainties can be reduced by adding additional particle labels or accounting for transmission efficiency, but coagulation will likely remain an inherent uncertainty. The limitations of transmission efficiency and coagulation are also noted at the end of the results section.

The caption in Figure 5 has been updated to state these factors more clearly

35: 6-10 now reads

"Notes (1) the soot in the blind mixture was known to be below the instrument detection limit and therefore is not expected to be found in the data, (2) coagulation of SOA and mineral dust, which occurred after aerosol input to the chamber, appears as the "other" category, (3) the aerosols types reported by AIDA do not account for PALMS transmission efficiency."

So my suggestion would be to split the paper and resubmit both parts in a thoroughly revised version one part with a clear focus on the algorithm and its general applicability to SPMS data including a real comparison to methods currently used (fuzzy-cmeans, manual decision tree, k-means). And the other with a thorough analysis of the blind data set explaining in a comprehensible way the measured spectra based on all available information, statistics and assumptions. If it is not possible to explain the measured spectra in a controlled laboratory experiment like the one described, the use of the instrument to characterize atmospheric aerosol populations would be quite limited.

The co-authors had a discussion regarding this suggestion but have decided that keeping the paper in something similar to the original format was the best course of action. Our reason is that the current format allows us to present a new technique for aerosol mass spectra classification and then use it on a single well constrained data set. We expect to expand its usage on future (e.g. field) data sets.

We acknowledge the reviewer's comments have focused on the general method and they raise an interesting point about using multiple method on a data set and then doing a cross comparison. We need to therefore comment that, in keeping this a new method to use with a demonstration on this data set, a multi-method comparison is well beyond the scope of our goal. The reviewer notes several different methods that have been used for aerosol mass spectral analysis. However, in practice, each instrumental group has focused on one or two techniques.

Setting up a means to apply several techniques to a single data set is therefore non-trivial and would occupy a significant amount of time. Again, we agree the cross comparison is interesting but would certainly represent months of research time beyond the scope of what we are attempting to do here.

Reviewer #2

As an overview comment we repeat the introductory review text here. Each point made is repeated specifically addressed below.

The major shortfall of this paper is that the authors neither explain the details of the machine learning approach fully nor do they fully engage with the aerosol classification results, leaving the reader somewhat confused in both realms.

In addition, the authors do not attempt to address the performance of their approach in terms of time or give information about how applicable it would be to ambient data sets where particles would not necessarily be of such distinct types.

Finally, they given no metrics for success – how good is good enough performance for this approach? How good are other methods, compared to that presented here?
I would recommend that this paper be significantly revised, in such a way that a) the machine learning approach can be fully described and choices made justified with data, and b) the aerosol particle classification results can be fully examined and compared to other methods.

Specific Comments:
1. The paper reads as if it was written by two separate people, one for the algorithm discussion and one for the aerosol particle classification discussion. This should be addressed as a final version (or versions) is developed. For example, on p. 4, the transition between lines 12 and 13 is abrupt and jarring.

We regret the paper seemed disjoint and believe this is partially from the fact the we are using a new technique on a more traditional dataset. It was not, as the reviewer suggested, written by two different people and knit together. We have gone through and tried to streamline the flow of the paper by removing the less pertinent details surrounding the algorithm, including your suggestion of simplifying the discussion of confusion matrices – please see the full track changes version. We are attempting to characterize two distinct topics as they related to the paper: The details surrounding training and applying a random forest as well as aerosol populations in the content of mass spectrometry.

To directly address this example, 4:12-17 now reads:

“To pick up on these minor yet important compositional differences, robust and generalizable analysis techniques are critical. We show that supervised training with random forests can differentiate aerosols in SPMS data more accurately than simpler approaches.”

2. The authors refer to “volatile” components of aerosol particles multiple times in the paper (first p. 3, line 13). I believe they mean semi-volatile, or at least “more volatile” than other components. Volatile species would not be expected to be found in particles.

The reviewer is correct and we have replaced the term volatile with semi-volatile.

3. In section 2.2, where the training data set is introduced, the authors need to discuss the applicability of this dataset to any "real" experiment. Would these particles be a good representation of ambient particles, for example?

We have addressed your question by including the following.
6: 6-11

> "The choice of supervised or unsupervised machine learning will depend on the researcher's use-case, and each method has unique advantages and disadvantages. We note a limitation of the random forest approach - and for supervised learning in general - is the inability to classify aerosol types outside of the training set. The ability of a random forest to characterize ambient atmospheric datasets, therefore, will strongly depend on which aerosols are contained within the training set."

Although it is feasible that unseen aerosol types will be assigned to the most chemically-similar label, supervised models are tuned to only make predictions on labels in the training set. The error statistics cannot be fully quantified for datasets with unknown aerosol types, so the model may not conform to the determined generalization error. In general, more particle types lead to a more generalizable classifier with better quantifiable error statistics. A study looking chiefly at atmospheric spectra would benefit from adding additional aerosol types and augmenting the analysis with existing methods such as clustering, which are designed to handle unlabeled data.

4. In the discussion of the data presented in Table 2, the authors state that the columns labeled "broad" are applicable to the categorization of the particles when they are lumped together into broad chemically-similar categories. These categories should be defined in this context (and not just in the context of the confusion matrices), presumably in Table 1. Any interpretation of the differences should be discussed in the results section. In addition, the paragraph on p. 11 about the 59+ ion observed in some samples (which ones?) should be moved into the results section. Finally, is it certain that 59+ is Co rather than an organic contaminant?

In this case, the contaminant is almost certainly Co^{+} originating from tungsten carbide grinder used to process some of the dust. A typical spectrum that shows the nature of the contaminant is shown below. The spectrum has markers that correspond to Co^{+}, W^{+2} and W^{+} but no obvious organic markers. Alternate assignments for m/z +59 are possible (and we present them in Table 2), but a prominent m/z +59 peak in this dataset is always associated with tungsten carbide contamination shown below.

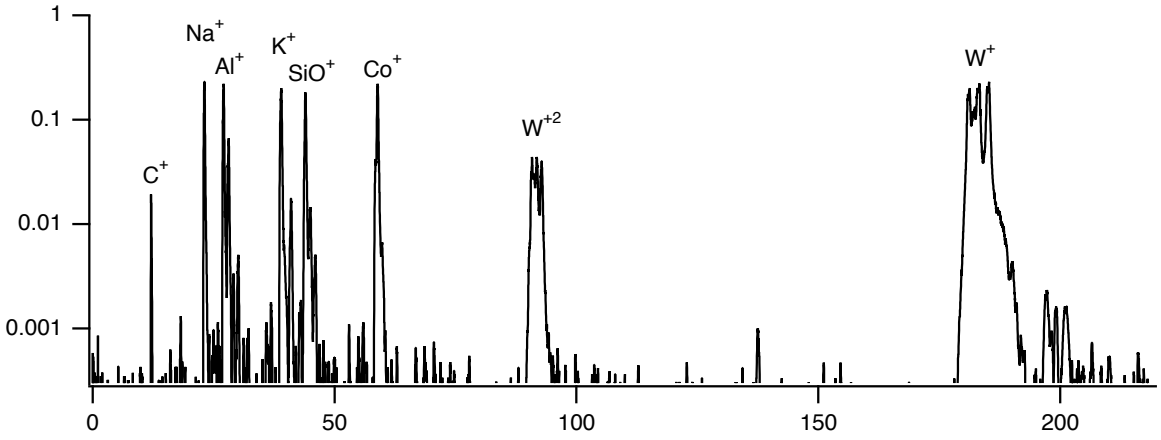


Table 1 has been color-coded to make the broad category definitions more clear. Additionally we move the paragraph into the discussion as requested and expand:

16:22 – 17:5

> "It is noteworthy that while most of the features are logical differentiators of the aerosol types investigated in FIN01 there were also surprises. One example is 59+ (cobalt), determined to be one of the most important features for differentiation. Further investigation determined this material was associated with tungsten carbide contaminant from dry powder dispersion equipment used on some samples. The contamination affected feldspar samples used during the second half of the AIDA measurements in particular."

5. The authors provide no information about the average mass spectra of the individual particle types and the variability within "identical" particles or between particle types. This would seem to be an important parameter in determining how well the algorithm can do to separate them. Based on the two peaks shown in Figure 1, this is an important factor.

To address your point, we have compared the method to a simple classifier which assigns unknown aerosols to the nearest class mean vector using the Euclidean distance metric. This answers your previous question of "how good is good enough" and provides a baseline that directly depends on the distance of aerosol mean vectors in feature space and variability of individual aerosols within each class. Figure 4 is updated to show the results of such a classifier. A couple of paragraphs have been included to summarize these results.
13:10 - 20 now reads:

> *"To access relative model performance, we contrast the results with a simple classifier that compares unseen aerosols to a set of class mean vectors. Using the Euclidean distance metric, the unknown aerosol is assigned to the nearest class. This simple baseline classifier helps put results in the context of machine learning techniques*

that rely on distance-based metrics such as k-means and hierarchical clustering. K-means clustering attempts to divide the data points into k distinct clusters, representing spectra as vectors. Using Euclidean distance, the standard algorithm assigns points to centroids, or clusters, which are essentially mean vectors representing the average of all points in the cluster. Assuming perfect convergence of k-means clustering, where k is the number of aerosol classes, each cluster represents the mean of aerosol in that class. The random forest results below demonstrate many areas of improvement over the simple classifier."

16:8 - 20 now reads

"*Across all categories, the random forest shows improvements over the Euclidean classifier in terms of both accuracy and precision. Figure 4 directly compares confusion matrices for the two methods, revealing overall accuracy improvements of at least 20%. The largest improvements are in the fertile soil and other category, where accuracy rises between 20% and 39% with the random forest. Computing the full confusion matrix for the Euclidean technique (as in figure 3) reveals similar results, with far more frequent mislabeling between fertile soils as well as coated/uncoated particles than our approach. These results reinforce the fact that chemically-similar aerosols which overlap in feature space will often be grouped together when using a single, distance-based classifier. The improvement from random forests is likely a result of a) the ensemble approach, which is known to produce better generalizability than single classifiers and b) the tendency of aerosols with similar chemical properties and atmospheric effect to appear mathematically distinct with a distance metric."*

6. The discussion of confusion matrices was confusing. Essentially, these matrices represent normalized counts of the sorting of known particles into the available classes.
This can be stated much more cleanly than the three page description provided on pp. 12 – 14. This is another example of a section that is trying to be both an algorithm and a particle chemistry paper, and not mixing the two effectively.

The section on confusion matrices has heavy revised and simplified to focus on aspects of the matrix that are directly used for the paper.

14: 2-4 reads:

"*A confusion matrix captures misclassification tendencies by pair-wise matching the model prediction with the true aerosol type or broad category [Powers, 2007], and can be understood as a contingency table matching model predictions to true labels."*

7. Figure 5 and the discussion of the "blind" test in Section 3.2 are key to the goals of this paper, but are confusing in their presentation. Regarding the text, why do the authors not know the number of particle types that were used in the challenge (p. 15, line 2 "3-4 aerosol types . . .were aerosolized")? How well can the results be evaluated if the test conditions are not known?...

The purpose of the blind experiment was to determine the capability of each mass spectrometer to determine the number of particle types and their composition; a situation deemed similar to the challenge of atmospheric sampling. This is now explicitly stated in the text at 17:10 “As part of the FIN01 workshop, it was known that an unknown number of aerosol types from Table 1 were aerosolized into the ADIA chamber at unknown size and relative concentration.” We realize the wording in the original sentence implied an unknown test but it was meant to indicate the participants were not aware.

b.) The authors describe a probabilistic correction for the mis-labeling that they observe in their confusion matrices (p. 15, lines 14 – 18), and say that the results “better represent the underlying aerosol population.” (p. 15, lines 17 – 18), but they don’t provide the data to evaluate this claim.

- The proposed probabilistic correction leads to insignificant changes in the final predictions as the computed fractional difference are small relative a) misclassifications between labels and b) uncertainties from having an unseen label. Furthermore, there is no guarantee the blind dataset will conform to the same mislabeling tendencies, as you have mentioned. We have removed the description from the paper and reapplied the method without the correction.

c.) The data presented in Figure 5 do not make the case that the authors are trying to make. While the two models (positive and negative) show relatively good agreement with each other, the representation of the particles introduced into the chamber is poor. The authors show the breakdown of soot, SOA, and mineral particles introduced in Figure 5, state that the soot particles are too small to see with their instrument, and then compare against the soot-containing dataset anyway. If the pie that represents “Aerosols Reported by AIDA” were renormalized to include only observable particles in this experiment, SOA would represent 44% of the pie and mineral would represent 56% – assuming that the “Aerosols Reported by AIDA” pie is also representing number of particles, rather than mass of introduced particles (this is not stated). If this pie represents something other than number, there is no comparison to the blind sample possible in this figure, only a comparison between the two models.

- The aerosols reported by AIDA are number concentrations, but there are several considerations to account for when considering the labels reported by our classifier in the blind dataset.
    - The aerosols reported by AIDA do not account for PALMS transmission efficiency, which depends on the size and aerodynamic properties of aerosols. For example, particles larger than 1000nm are over-reported by the classifier due to increased PALMS efficiency in that range. We now note this inherent limitation.
    - During the course of the experiment, we expect (and observed) the mineral dust and SOA to coagulate. Since aerosol types were reported by AIDA before particles enter the chamber, it is not possible to quantify exactly what fraction of the particles picked up an SOA coating. Additionally there is the possibility of

effectively producing a particle type not in the training set, depending on the exact mineralogy of the mineral dust used by AIDA. While it is known a mineral dust was included in the chamber, the exact composition of the dust is not know. The mineral may either contain a specific component or a soil dust. While our training set contains K-Feldspar coated with SOA, a different type of SOA-coated mineral dust will appear unique to the model. Because the generalization performance of supervised classifiers is ill-defined for particles not included in the training set, this could lead to performance that is not captured by the confusion matrices. Given the experimental uncertainties from transmission efficiency and coagulation, as well as the model uncertainties highlighted in the confusion matrices, we believe the results reveal skill in using random forests to pick out distinct aerosols. In future studies, uncertainties can be reduced by adding additional particle labels or accounting for transmission efficiency, but coagulation will likely remain an inherent uncertainty.

These are stated at the end of our results section as well as the caption for Figure 5. The caption in Figure 5 has been updated to mention transmission efficiency as an uncertainty.

35: 6-10 now reads

"Notes (1) the soot in the blind mixture was known to be below the instrument detection limit and therefore is not expected to be found in the data, (2) coagulation of SOA and mineral dust, which occurred after aerosol input to the chamber, was often categorized as mixed mineral and organic particles or fertile soils (i.e., mixtures of mineral and organic components) considered in the training data set, (3) the aerosols types reported by AIDA do not account for PALMS transmission efficiency (see text for details). "

Technical Corrections:

1. P. 10, line 17: remove the word "rows" from the line.

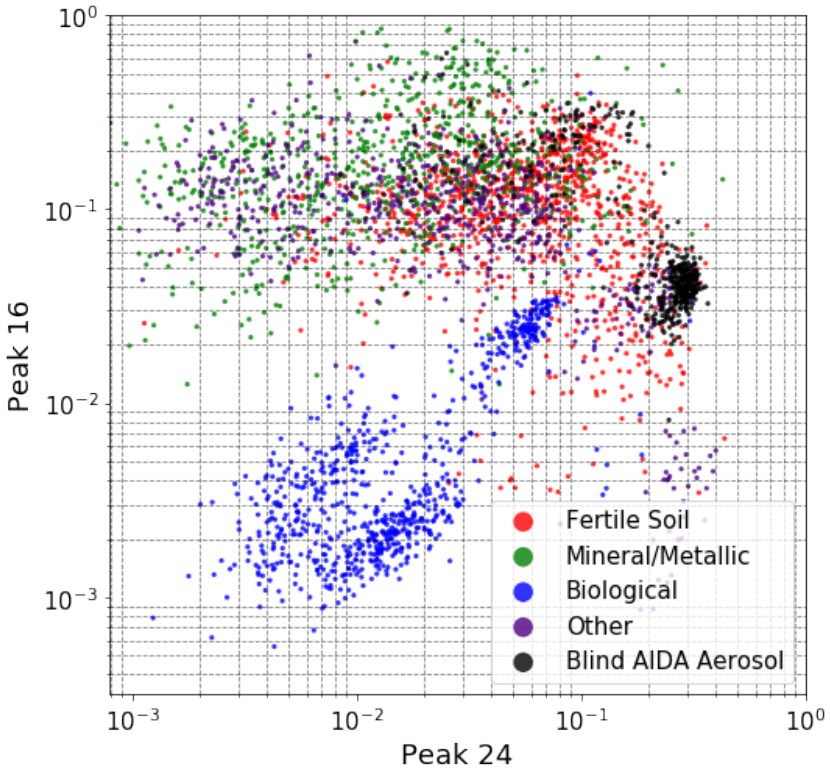
This has been corrected.

2. P.15, line 3: "AIDA" is written as "ADIA."

This has been corrected.

3. Figure 1 is very difficult to read. The black points for the "Blind AIDA Aerosol" are only visible on the right-most part of the graph, in the region of (0.3, 0.04) and the similar colors are hard to differentiate. Consider a figure like this that is broken out into the broad categories of particle types.

As suggested, we created a similar scatter plot using only broad categories. We still find it difficult to pick out each of the particle types in the region in (0.3, 0.4), even when plotting only small subsets of the training set. This is a result of significant overlapping of aerosol types in the region, and difficult to alleviate. Nevertheless, the aim of the figure is to demonstrate that some types overlap significantly, while others such as SOA form distinct clusters. Since both versions of the plot demonstrate this to a similar degree, we have decided to leave the original figure in the paper.



4. Figure 5, negative model, the small pie includes 5 wedges but only 4 labels.
This has been corrected.

A Machine Learning Approach to Aerosol Classification for Single Particle Mass Spectrometry

Christopoulos, Costa D.[1], Garimella, Sarvesh[1,2], Zawadowicz, Maria A.[1,3], Möhler, Ottmar[4] and Cziczo, Daniel J.[1,5]

[1] Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA, United States

[2] ~~now at~~ ACME AtronOmatic, LLC, Portland, OR, United States

[3] ~~now at~~ Atmospheric Sciences and Global Change Division, Pacific Northwest National Laboratory, Richland, WA, United States

[4] Institute of Meteorology and Climate Research, Karlsruhe Institute of Technology, Karlsruhe, Germany

[5] Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, United States

Abstract

Compositional analysis of atmospheric and laboratory aerosols is often conducted via single-particle mass spectrometry (SPMS), an *in situ* and real-time analytical technique that produces mass spectra on a single particle basis. In this study, machine learning classifiers are created using a dataset of SPMS spectra to automatically differentiate particles on the basis of chemistry and size. Machine learning algorithms build a predictive model from a training set for which the aerosol type associated with each mass spectrum is known *a priori*. ~~Our primary focus surrounds the growing of random forests using feature selection to reduce dimensionality, and the evaluation of trained models with confusion matrices. In addition to classifying ~20 unique, but chemically similar, aerosol types,~~Classification models were also created to differentiate aerosol within four ~~broader~~broad categories: fertile soils, mineral/metallic particles, biological, and all other aerosols. Differentiation was accomplished using ~40 positive and negative spectral features. For the broad categorization, machine learning resulted in a classification accuracy of ~93%. Classification of aerosols by specific type resulted in a classification accuracy of ~87%. The 'trained' model was then applied to a 'blind' mixture of aerosols which was known to to be a subset of the training set. Model agreement was found on the presence of secondary organic aerosol, coated and uncoated mineral dust and fertile soil.

1. Introduction

~~Following the introduction of random forests in the 1990s, recent developments in deep learning and neural networks have triggered a renewed interest in machine learning. This has led to the development of numerous easy-to-use, freely-available, open-source packages in popular programming languages like Python, and these tools are becoming increasing used in academia and industry. While random forests have been used for complex classification and regression analysis in various fields, studies that employ random forests in aerosol mass spectrometry remain sparse. Utilizing these tools, the primary purpose of our study is to introduce a framework for growing random forests, reducing dimensionality, ranking chemical features, and evaluating performance using confusion matrices. Such properties are desirable for SPMS studies, where input variables can become redundant and interpretability is more limited with methods such as cluster analysis and neural networks. Analysis techniques such as those falling out of recent artificial intelligence research can prove useful for helping to tease out the subtle yet significant impact that aerosol chemistry has on the climate system.~~

~~Atmospheric aerosols impact clouds and the Earth's radiative budget. A lack of understanding of aerosol composition therefore contributes to uncertainty in determination~~ The interaction of atmospheric aerosols with clouds and radiation contributes to the uncertainty in determinations of both anthropogenic and natural climate forcing [Boucher et al., 2013; Lohmann and Feichter, 2005]. Aerosols directly affect atmospheric radiation by scattering and absorption of radiation from both solar and terrestrial sources. The radiative forcing from particulates in the atmosphere depends on optical properties that vary significantly among different aerosol types [Lesins et al., 2002]. Aerosols also indirectly affect climate via their role in the development and

maintenance of clouds [Vogelmann et al., 2012; Lubin et al., 2006]. Ultimately, the formation, appearance, and lifetime of clouds are sensitive to aerosol properties like shape, chemistry, and morphology [Lohmann and Feichter, 2008]. Characterization of aerosol properties, therefore, plays a vital role in understanding weather and climate.

The chemical composition and size of aerosols has been analyzed on a single particle basis *in situ* and in real-time using single particle mass spectrometry (SPMS; Murphy [2007]). First developed ~2 decades ago, SPMS permits the analysis of aerosol particles in the ~150 – 3000 nm size range, while differentiating internal and external aerosol mixtures and characterizing both ~~semi-~~volatile (e.g. organics and sulfates) and refractory (e.g. crystalline salts, elemental carbon and mineral dusts) particle components. Particles are typically desorbed and ionized with a UV laser and resultant ions are detected using time-of-flight mass spectrometry [Murphy, 2007]. A complete mass spectrum of chemical components is normally produced from each analyzed aerosol particle [Coe et al., 2006]. Despite almost universal detection of components found in atmospheric aerosols, SPMS is not normally considered quantitative without specific laboratory calibration [Cziczo et al., 2001].

~~Chemical composition of an individual atmospheric aerosol particle is a complex interplay between its primary composition at the source (i.e. dust, biogenic organic, anthropogenic organic, soot, etc.) and its atmospheric processing up to the time of detection. Atmospheric processing can include a combination of coating~~Aerosols with ~~secondary material, coagulation and cloud processing. Even~~ different ~~primary aerosol types~~properties can ~~have~~appear similar ~~mass spectral markers.~~in the context of SPMS. For example, fly ash~~,~~ and mineral dust ~~and bioaerosol can all~~ contain ~~strong phosphate~~

Formatted: Font color: Black

Formatted: Font color: Black

Formatted: Font color: Black

~~signal~~peaks corresponding to silicates, phosphates, metals, and metal oxides despite different origins and emission sources [Zawadowicz et al., 2017]. ~~Secondary material is often difficult~~This complicates analysis of aerosol populations because their properties need to ~~differentiate from primary material, but even~~be well-defined in order to increase agreement between models and observations [Niemand et al., 2012; Hoose and Möhler, 2012; Welti et al., 2009]. Even minor compositional changes can be atmospherically important. As one example, mineral dusts are known to be effective at nucleating ice clouds~~; however, despite~~ [Cziczo et al., 2013]. Particles in the atmosphere undergo chemical and morphological changes as they age and eventually contain material from several sources [Boucher et at. 2013]. Despite minor addition of mass, ~~atmospherically processed~~aged mineral dust is less suitable for ice formation [Cziczo et al., ~~2013].~~2013], but these particles then act as cloud condensation nuclei and participate in warm cloud formation [Andreae et al., 2008]. As a second example, ice nucleation in mixed-phase clouds has been suggested to be predominantly influenced by feldspar, a single component among the diverse mineralogy of atmospheric dust [Atkinson et al., 2013]. ~~Using current SPMS data analysis approaches, it is difficult to detect these minor yet important compositional differences and new robust and generalizable analysis techniques are critical.~~

~~We~~Here we show that supervised training ~~with random forests~~and a rule-based probabilistic classification of a decision tree ensemble can ~~differentiate aerosols in~~be used for differentiation of SPMS ~~data more accurately than simpler approaches~~spectra. Various clustering methods have been used to group aerosol types [Murphy et al., 2003; Gross et al., 2008] but these algorithms are known to struggle with chemically-similar

Formatted: Font color: Black

aerosols as they do not incorporate known particle labels in the training process. Such 'unsupervised' clustering algorithms automatically group unlabeled data points on the basis of a specified distance metric in feature space, in this case mass spectral signals. For the purposes of setting broad aerosol categories, which are chemically similar and easily separable in feature space, clustering is the simpler tool and the data easier to interpret. For identifying new or potentially unexpected atmospheric aerosols, such properties are desirable; however, the advantages of clustering greatly diminish when considering similar particle types that overlap in feature space. ~~A limitation often encountered is the need to manually reduce the number of final clusters due to grouping of mathematically similar yet chemically distinct aerosols [Murphy et al 2003].~~ Fertile soils, for instance, are often grouped into a single category despite different sources and atmospheric histories. Clustering algorithms should therefore be considered as a tool to use alongside supervised classification. The latter may be used to further explore unique aerosol types or verify manually labeled clusters with higher precision. Furthermore, the ensemble approach presented here also produces ~~interpretable~~ variable rankings and probabilistic predictions that assist in addressing measurement uncertainty. ~~The choice of supervised or unsupervised machine learning will depend on the researcher's use-case, and each method has unique advantages and disadvantages. We note a limitation of the random forest approach – and for supervised learning in general – is the inability to classify aerosol types outside of the training set. The ability of a random forest to characterize ambient atmospheric datasets, therefore, will strongly depend on which aerosols are contained within the training set.~~

In this study, we demonstrate the capabilities of ~~random forests~~machine learning

to automatically differentiate particles on the basis of chemistry and size. The resulting model can capture minor compositional differences between aerosol mass spectra. By testing predictions using an independent, or 'blind', dataset, we illustrate the feasibility of combining on-line analysis techniques such as SPMS with machine learning to infer the behavior and origin of aerosols in the laboratory and atmosphere.

2. Methodologies

2.1 PALMS

The Particle Analysis by Laser Mass Spectrometry (PALMS) instrument was employed for these studies. PALMS has been described in detail previously [Cziczo et al. 2006]. Briefly, the instrument samples aerosol particles in the size range from ~200 to ~3000 nm using an aerodynamic lens inlet into a differentially-pumped vacuum region. Particle aerodynamic size is acquired by measuring particle transit time between two 532 nm continuous wave neodymium-doped yttrium aluminum garnet (Nd:YAG) laser beams. A pulsed UV 193 nm excimer laser is used to desorb and ionize the particles and the resulting ions are extracted using a unipolar time-of-flight mass spectrometer. The resulting mass spectra correspond to single particles. The UV ionization extracts both refractory and ~~semi-~~volatile components and allows analysis of all chemical components present in atmospheric aerosol particles [Cziczo et al. 2013].

2.2 Dataset

A set of 'training data' was acquired by sampling atmospherically-relevant aerosols. The majority of the dataset was acquired at the Karlsruhe Institute of Technology (KIT) Aerosol Interactions and Dynamics in the Atmosphere (AIDA) facility

during the Fifth Ice Nucleation workshop — Part 1 (FIN01). The remainder were acquired at our Aerosol and Cloud Laboratory at MIT. The FIN01 workshop was an intercomparison effort of ~10 SPMS instruments, including PALMS. The training data correspond to spectra of known particle types that were aerosolized into KIT's main AIDA and a connected auxiliary chamber for sampling by PALMS and the other SPMSs (Table 1). Hereafter we group both chambers with the name 'AIDA'. The number of training spectra acquired varied by particle type, ranging from ~250 for secondary organic aerosol (SOA) to ~1500 for potassium-rich feldspar ("K-feldspar"). In total, ~50,000 spectra are considered with each spectrum containing 512 possible mass peaks and an aerodynamic size. (Table 2). Additionally, the FIN01 workshop included a blind sampling period, where AIDA was filled with ~~an unknown number of~~3 - 4 aerosol types known to be from the training set (i.e., for which spectra had already been acquired) but (*a priori*) of unknown size, specific types and at unknown concentrations.

Figure 1 illustrates a simple differentiation of particles using only two mass peaks in one (negative) polarity. Mass peaks represent fractional ion abundance, measured as a normalized total signal (ion current). In this example, the normalized areas of negative mass peaks 24 (C_2^-) and 16 (O^-) are plotted. Distinct aerosol types are differentiated by color with clusters forming in this two-dimensional space. Note that spectra of the same aerosol type form distinct clusters (e.g. Arizona Test Dust, ATD), as do similar aerosol classes (e.g., soil dusts). Co-plotted in Figure 1 are data from the blind experiment. Distinct clusters of spectra from the blind experiment are noticeable and correlate with known clusters. Described in the next section, machine learning algorithms draw "decision boundaries" that best separate different groups of data points based on set of

rules. Machine learning is not bound by the simplistic two- dimensional space shown in Figure 1 and instead uses all 512 mass peaks and aerodynamic size.

2.3 Aerosol Classification

A trained classification model maps a continuous input vector 'X' to a discreet output value using a set of parameters 'learned' from the data. Figure 2 illustrates the mapping of a mass spectrum to vector space. In contrast to traditional, hard-coded, rule-based classification methods, machine learning determines parameters that partition the data set. To form X, mass spectra are converted to dimensional vectors normalized to the total ion current (i.e., the total of all mass peaks sum to 1 in each spectrum). The elements of the vectorized mass spectrum, termed 'features', hold information about the ionization efficiency and relative abundance of chemical species in each aerosol and serve as the variables for the machine learning model.

Machine learning is conducted in two phases: training and testing. During training, a model is constructed and iteratively updated based on data (i.e., mass spectra) from the training set. For this work, the set of known aerosol types sampled by PALMS was converted to dimensional vectors. These data form the basis set for defining each aerosol type. ~~A random forest~~An ensemble of decision trees was used to generate predictions of aerosol type. A single decision tree is a statistical decision model that performs classification based on a series of comparisons relating a variable X_i (in this case a normalized mass peak in X) to a learned threshold value [Breiman, 2001]. Represented as an algorithmic tree, a binary decision tree consists of a hierarchy of nodes where each node connects via branches to two other nodes deeper in the tree. At each node, one of the two branches is taken based on whether a normalized peak X_i is greater or less than a

threshold value. Each branch leads to another node where a different test is performed. After a series of tests, one at each node, a class is assigned to a given sample; these are the so-called 'leaves'. Figure 2 illustrates the classification model for a single decision tree.

Each test in the tree narrows the set of reachable output leaves and thus the sample space of possible aerosol labels. After *h* tests in this study, where *h* ranges from 10 to 3000, the set of reachable leaves and possible labels is 1 and the decision tree outputs a prediction. Because PALMS is unipolar – either a positive or negative mass spectrum is produced – simultaneous generation of positive and negative spectra on a particle-by-particle basis is not possible. Two separate classification models, one for each polarity, were therefore generated to classify aerosols. These are hereafter referred to as the 'positive' and 'negative classification algorithms'.

2.4 ~~Random Forests~~Decision Tree Ensembles

~~A random forest is an~~An ensemble consists of ~~decision tree~~ a collection of classifiers where each ~~classifer~~ independently labels ~~an unknown~~a spectrum vector X. To make a final prediction of aerosol type, decision trees within an ensemble 'vote' on a classification label. Each vote has equal weight and the spectrum is assigned to the majority choice. Each tree within an ensemble is independently grown on a subset of the training data so that a commonly voted label implies a higher certainty. Adding members to an ensemble increases the robustness of a classification model by providing alternative hypotheses and is therefore preferable to single classifiers.

Before an ensemble method is implemented for classification, trees are independently grown during training. A total of k trees, with k = ~~110~~1000, were grown using a bootstrap sample from the training set. In bootstrap sampling, each tree sees an independent sample set of equal size drawn from the full training set by sampling spectra with replacement. On average, each tree is built with ~63% of the ~~original~~ data~~, leaving a portion of the training set unsampled.~~ The unsampled data ~~for each tree~~, known as 'out-of-bag' observations, ~~are recorded and later~~ provide a means to assess classification error for ~~the forest. To determine model error, predictions are made for each point in the dataset using only the subset of trees that did not use the point for training. Each training point is left out at least once. This is analogous to making predictions with a separately trained forest that did not observe the point and prevents testing with the same data used for training.~~ each tree during the training process.

Given a bootstrap sample, a binary decision tree is grown by sequentially creating tests that maximize the separation between classes in parameter space. A test is created by defining a comparison that minimizes the information entropy of a possible split, thus minimizing the randomness of prediction labels [Breiman, 1996]. To generate variability in the model, a best split is chosen among a random set of possible splits at each node on the basis of entropy [Breiman, 2001]. After iteratively defining thresholds for each new node, the tree grows in size until a series of tests ending at some node S_q uniquely characterizes an aerosol as a particle type. A leaf is then appended to node S_q with the corresponding label. In classification mode, an aerosol spectrum that passes the same tree will undergo the same series of tests and will end in the same leaf, thus being labeled in the same way. For the purposes of this study, each tree had ~3,300 nodes.

~~The number of variables per split is chosen to be 11 and the number of trees is 110. The optimal model was determined by enumerating combinations of these parameters on a coarse grid and selecting the values that produce the lowest test error. Model behavior is primarily sensitive to the number of variables per split, and shows weak dependence on the number of trees and number of input variables beyond small values. As the number of variable splits increases, error decreases exponentially to a local minimum before again rising due to over fitting. Alternatively, as the number of trees is increased the error asymptotes to some nonzero value, a known characteristic of random forests where test error converges to the generalization error. The models were trained with the Python 2.7 Scikit-learn module on a MacBook Pro with 16 GB 1600 MHz DDR3 memory and a 2.5 GHz Intel Core i7 processor. A typical random forest model took about 5-10 seconds to train, and we found a linear relationship between runtime and both the number of trees and variables per split.~~

~~Overall, the generalizability and robust performance of random forests is owed significantly to the series of random statistical procedures used to construct such models. An ensemble classifier reduces variability by averaging predictions over a series of independently trained models, and bagging introduces additional randomness by producing "perturbed" versions of the original data via random sampling of input data. The randomness used in constructing forests, both in bagging the training set and choosing variable splits, work to decorrelate the output of each tree even as the inputs become correlated [Breiman, 2001]. As the number of trees increases, the law of large numbers guarantees a convergence of the out-of-bag error to the generalization error.~~

2.5 Dimensionality Reduction and Chemical Feature Selection

Dimensionality reduction is the process of representing data with fewer variables than initially present in the dataset, in this case less than the original 512 mass peaks and aerodynamic size. In addition to facilitating data visualization, reducing computation time and limiting overfitting [Mjolsnes, 2001], dimensionality reduction, in the context of aerosol mass spectra, also indicates the most important chemical makers for differentiation. Feature ranking was algorithmically determined by comparing the performance of trees before and after removing information about peak X_i. The method is that the values of variable X_i is permuted for tree *k* in the out-of-bag set so that the variable is irrelevant to the final label. The change in misclassification before and after the permutation is calculated and then repeated for all trees so that a variable ranking is obtained [Breimann, 2001]. Table 2 rows ranks mass peaks (features) by polarity in importance using this method. The columns at left list feature rankings (i.e., most to least important for correct classification) for the entire set of aerosol types. The columns at right list rankings when aerosol types are grouped into the broad, chemically similar, categories. A final ranking was determined by sequentially adding variables and observing classification performance response. All variables preceding two e-foldings in classification error were maintained in the final model. Both the specific aerosol type and broad aerosol category models were retrained using this subset of the initial variables, listed in Table 2.

~~2.5 Comparison to Euclidean Distance Classifier~~

~~To access relative model performance, we contrast the results with a simple classifier that compares unseen aerosols to a set of class mean vectors. Using the Euclidean distance metric, the unknown aerosol is assigned to the nearest class. This~~

~~simple baseline classifier helps put results in the context of machine learning techniques that rely on distance-based metrics such as k-means and hierarchical clustering. K-means clustering attempts to divide the data points into k distinct clusters, representing spectra as vectors. Using Euclidean distance, the standard algorithm assigns points to centroids, or clusters, which are essentially mean vectors representing the average of all points in the cluster. Assuming perfect convergence of k-means clustering, where k is the number of aerosol classes, each cluster represents the mean of aerosol in that class. The random forest results below demonstrate many areas of improvement over the simple classifier.~~

It is noteworthy that while most of the features are logical differentiators of the aerosol types investigated in FIN01 there were also surprises. One example is 59^{+} (cobalt), determined to be one of the most important features for differentiation. Further investigation determined this material was a contaminant from dry powder dispersion equipment used on some samples. This serves to illustrate the lack of *a priori* judgment by the algorithm and an unintended benefit of machine learning process (i.e., contamination identification).

3. Results

3.1 Confusion Matrices and Probabilistic Model Performance

A confusion matrix captures misclassification tendencies by pair-wise matching the model prediction with the true aerosol type or broad category [Powers, 2007~~], and can be understood as a contingency table matching model predictions to true labels.~~].

Confusion matrices represent model predictions as columns i and true aerosol type of category as rows j, where class names are mapped to integers i , $j \in \{1,2, \dots , y\}$. In this study, matrices have been normalized along each column to show the fraction of aerosols labeled as j that actually belong to i (Figures 3 and 4). For aerosol classification, these matrices can also be interpreted as similarity measures between particle types. Since the basis of decision tree classification is mathematical separation of physical quantities, misclassifications result from similarity in mass peaks and their ion abundance between aerosol types. This is most easily visualized as overlapping clusters in the simple two dimensional space in Figure 1.

Because the size of the set is large (~22,300), the general classification behavior can be quantified in term of conditional probability. If $\hat{Y}_i$ is the set of predicted aerosol spectra with aerosol label i and Y_j is the corresponding set of true spectrum-label pairs for label j, then the conditional probability of assigning an aerosol to label i given a predicted label j is given by:

$$p(i \mid j) = \frac{|Yj \cap \hat{Y}i|}{|Yj|} \quad (1)$$

C is the raw confusion matrix of spectrum counts and $p(i \mid j)$ is the conditional probability distribution over all true aerosol labels i, conditioned on some model-generated label j. To obtain matrix P, which encodes $p(i \mid j)$ for all possible labeling, columns of C are normalized with respect to the total aerosol counts for each label with Eq. 1.

Model performance for each aerosol is summarized in the diagonal elements of ~~the confusion matrix~~P, which represent the fraction of aerosol in column j labeled correctly. The classification accuracy (a) is given by averaging diagonal elements of P. A

perfect classification model produces the identity matrix, as all data points are classified correctly 100% of the time. For example, in the positive confusion matrix, SOA and Agar growth medium are correctly labeled in the test set 100% of the time. Barring element truncation, all columns of P add to 1.

Figures 3 and 4 display confusion matrices as heat maps for the full set of particle labels and broad grouped particle categories, respectively. Broad categories are delineated by bold horizontal and vertical lines in Figure 3 as fertile soil (Argentinian, Chinese, Ethiopian, Moroccan and two German soils), pure mineral dust and metallic particles (ATD, illite NX, fly ash, Na-feldspar, K-feldspar), biological (Agar growth medium, *P. syringae* bacteria, cellulose, Snomax, and hazelnut pollen), and other (K-feldspar with sulfuric acid (SA) and SOA coatings, soot, and SOA) particles. Some model confusion exists between fertile soils and coated/uncoated feldspars which can be explained since soils are mineral dust mixed with organic and other materials.

Positive mass spectra appear to hold more information with respect to differentiating aerosols than negative. Label-wise classification accuracy for the negative algorithm ranges from 3-5% lower. A large part of this performance discrepancy is due to greater ability of positive spectra to differentiate coated particles within the 'other' category.

In addition to quantifying misclassification tendencies between classes, the confusion matrix can be redefined to show confusion for aerosols within broad categories themselves. Intraclass misclassification analysis is accomplished by considering smaller portions of C and using the same probabilistic assumptions highlighted for the full confusion matrices to form modified probability distributions. The full confusion matrix

is partitioned into submatrices representing confusion in a specific aerosol category and renormalized with respect to matrix columns. L is the subset of particle labels of a broader set of aerosols. Integrating the full conditional probability distribution over labels that are impossible to observe gives the probability distribution over members of L:

$$P_l(i,j) = p(i \in L \mid j \in L) = \frac{C(i \in L, j \in L)}{\sum_{i' \in L} C(i', j \in L)} \qquad (2)$$

For example, to determine $P_l(i \mid j)$ for fertile soils, a submatrix is formed by collecting spectral counts in the first 6 rows and columns of the full confusion matrix (Figure 3). Column normalization is then applied to derive a probability distribution over labels in the fertile soil category, conditioned on the aerosol actually being a fertile soil. This analysis is repeated over all categories in both models. Finally, the relative performance of both models is isolated and considered with respect to each specific aerosol category.

The precision score [Powers, 2007] captures the classification behavior for some subset of aerosol L by averaging fractions of correctly classified aerosols for labels within that category:

$$\text{Precision Score(L)} = \frac{1}{|L|} \sum_{i=j}^{|L|} P(i \in L, j \in L) \qquad (3)$$

When applied to P_l, the precision score captures classification performance on a population with only aerosol labels contained in L. The algorithm is expected to correctly label an aerosol in such a population with a probability equal to the precision score. The precision score is valuable when using the classification model as a particle screener, producing probability distributions over a subset of aerosol labels of interest. The

confusion characteristics are shown in Table 3 for each category in terms of the precision score and the mean and standard deviation of misclassification within each category. Although both models perform similarly for biological spectra, discrepancies of 2-5% appear in the remaining categories. For regimes consisting of only mineral/metallic or other particles, the positive algorithm shows intraclass performance advantages in terms of the precision score, but most notably in terms of fewer mislabeling of mineral/metallic particles. The largest precision discrepancy is observed for fertile soils, where the positive ion algorithm has a 5% advantage in precision with approximately half the false labeling rate.

~~Across all categories, the random forest shows improvements over the Euclidean classifier in terms of both accuracy and precision. Figure 4 directly compares confusion matrices for the two methods, revealing overall accuracy improvements of at least 20%. The largest improvements are in the fertile soil and other category, where accuracy rises between 20% and 39% with the random forest. Computing the full confusion matrix for the Euclidean technique (as in figure 3) reveals similar results, with far more frequent mislabeling between fertile soils as well as coated/uncoated particles than our approach. These results reinforce the fact that chemically-similar aerosols which overlap in feature space will often be grouped together when using a single, distance-based classifier. The improvement from random forests is likely a result of a) the ensemble approach, which is known to produce better generalizability than single classifiers and b) the tendency of aerosols with similar chemical properties and atmospheric effect to appear mathematically distinct with a distance metric.~~

~~Beyond classification, the obtained variable rankings alone provide interesting~~

~~insights into the dataset. It is noteworthy that while most of the features are logical differentiators of the aerosol types investigated in FIN01 there were also surprises. One example is 59^{+} (cobalt), determined to be one of the most important features for differentiation. Further investigation determined this material was associated with tungsten carbide contaminant from dry powder dispersion equipment used on some samples. The contamination affected feldspar samples used during the second half of the AIDA measurements in particular. This serves to illustrate the lack of *a priori* judgment by the algorithm and an unintended benefit of machine learning process (i.e., contamination identification).~~

3.2 Characterization of Blind Data

As part of the FIN01 workshop, it was known that ~~an unknown number of~~3 - 4 aerosol types from Table 1 were aerosolized into the ADIA chamber but at unknown size and relative concentration. PALMS, one member of the blind intercomparison effort, collected ~25,000 spectra. After data analysis, the aerosol types and relative abundances were provided to each group (Figure 5, top center).

The presence or absence of particle types in the blind set was initially diagnosed by choosing particles predicted at or above the 1% level. We note here that this step was based on the knowledge that (1) a distinct set of particles would be placed in the chamber and (2) particles present at or below the 1% level were most likely contamination. We further note that this step is unique to a blind study and would not be applicable to the atmosphere. Normalized confusion matrices were redefined for the aerosols in the population (i.e., those above the 1% level), which forms the labels of set L in Eq. 2.

Formatted: Indent: First line: 1.27 cm, No widow/orphan control, Don't adjust space between Latin and Asian text, Don't adjust space between Asian text and numbers

Finally, particle counts are re-computed by reassigning particle labels based on the modified confusion matrix. For each particle label *j*, a fraction $n' = P(i \mid j)$ of particles labeled as *j* are reassigned to *i*. This probabilistic correction accounts for aerosol mislabeling tendencies observed during testing, producing statistics that better represent the underlying aerosol population. The expected fraction of particles belonging to label *i* (denoted $\hat{n}i$) is given by:

$$< \hat{n}i > = \frac{< n_i >}{|n|} = \frac{1}{|n|} \sum_j P(i \mid j) \, |n_j| \quad (4)$$

where *n* is a set containing all blind spectra and n_j is the set of particles labeled as *j*.

Figure 5 illustrates the results after this step, where the bottom charts show corrected fractional percentages for each aerosol category. Because SOA was nearly always labeled correctly (Figure 3), the remaining aerosols are considered separately using the full set of candidate aerosol labels. Both positive and negative models arrived at similar results, with inconsistencies primarily associated with the presence of trace fertile soils and mineral dust / fly ash particles. The positive algorithm identifies ~2-4% of the AIDA population as each Argentinean soil, German soil, ATD, and cellulose whereas the frequency of these aerosols was too low to consider in the negative. Alternatively, the negative model estimates Na-Feldspar at ~~14~~8% of the total population, a label not identified by the positive algorithm. This discrepancy can be explained by the 1% selection criterion for aerosols present in the population. Fertile soils, ATD, and cellulose frequently accumulate error along rows in the full positive confusion matrix, indicating frequent confusion with other categories (Figure 3). Furthermore, with the observed misclassification rates ranging ~1-4%, it is expected that these aerosol labels are false positives. The negative model offers an alternative hypothesis, suggesting these

Formatted: Font: Arial Unicode MS, 13 pt

miscellaneous aerosols are Na-feldspar. Since there is significant model agreement on the percentages of SOA, K-Feldspar, and coated feldspars, this part of the blind mixture population (~90%) can be characterized with most certainty. For the disputed aerosol labels, more credence is lent to the negative classification algorithm on the basis of improved precision for fertile soils.

The aerosols reported in the blind mixture were soot, mineral dust, and SOA. This mineral component was not defined and may have been either a specific mineral or soil dust. The soot aerosols were below the cutoff diameter for PALMS; they were therefore not detected or identified by the algorithms. Similarly, particles with diameters greater than ~1000 nm are detected with increasingly large inefficiency which likely leads to undercounting of mineral dust [Cziczo et al., 2006]. Both algorithms robustly labeled SOA with large agreement, consistent with the 100% accuracy observed in the test set.

SOA coated mineral dust was identified as a particle type. This material was not directly input to AIDA but the report is most likely correct, due to coagulation within the AIDA chamber during the course of the blind experiment. ~~This may also explain some indications of fertile soils, which are known to be mixtures of mineral and organic components.~~ The training data set did not contain coagulated SOA and mineral dust but did include SOA-coated K-Feldspar, which explains the identification.

While both models identified a variety of fertile soils, and not a single type, these results are largely consistent with the ~~presence of coagulated organics and minerals and the~~ known uncertainties highlighted by the confusion matrices discussed previously. Given the presence of any single mineral dust, some confusion with fertile soils, SA coated Feldspar, and Na-Feldspar is expected (Figure 3). Moreover, as discussed

previously [Gallavardin et al., 2008], AIDA backgrounds are not completely particle-free. During the FIN01 study, contamination particles from previous test aerosol were frequently observed as background and they could also be the origin of some low-concentration particles matching fertile soil chemistry.

Formatted: Suppress line numbers

4. Conclusions and Future Work

~~This study lays out a framework for training and implementing random forests on SPMS data, with a focus on dimensionality reduction and the evaluation of model performance with confusion matrices. A key benefit to the proposed method is chemical feature selection, which allows researchers to identify potentially important chemical markers between arbitrary groups of aerosols or identify sources of contamination. Additionally, the approach~~The machine learning approach described here allows for differentiation of aerosols within a SPMS dataset, augmenting existing tools and reducing the need for a qualitative comparison between mass spectra. This study lays out a framework for training and implementing an ensemble classification model and interpreting results in the context of laboratory and atmospheric aerosol populations. Across a representative sample of possible aerosol types, the behavior of each algorithm predictably allows users to infer the presence or absence of specific aerosols and quantify aerosol abundance. Machine learning is automated and the output of the model must then be informed by human knowledge of aerosol chemistry. Machine learning should therefore be considered as an additional tool to interpret mass spectra to better distinguish aerosols with unique properties in terms of atmospheric chemistry, biogenic cycles, and

population health.

The ~~random forest~~ensemble decision tree classification framework described here may be generalized to any instrument, or set of instruments, capable of collecting physical and chemical information that distinguishes particles. Although the method described here is applied to a stand-alone SPMS and tested with a set of 'blind' data, ancillary laboratory or field data can be integrated to expand the data set. The success of these algorithms is data-dependent, where better performance is expected for instruments that provide more, and more quantitative, analysis of the aerosol properties. Although the algorithms implemented in this study were primarily used to categorize SOA, mineral dust, fertile soil and biological aerosols, these models can adopt an arbitrary large set of aerosol data.

Acknowledgements

We thank the FIN01 and AIDA teams for logistical support and scientific discussions. We acknowledge funding from NSF which allowed our participation (grant AGS-1461347). M.A.Z. acknowledges the support of NASA Earth and Space Science Fellowship and D.J.C. acknowledges the support of Victor P. Starr Career Development Chair.

References

Andreae, M. & Rosenfeld, D.: Aerosol–cloud–precipitation interactions. Part 1. The nature and sources of cloud-active aerosols, Earth-Sci. Rev., 89, 13-41, doi:10.1016/j.earscirev.2008.03.001, 2008.

Atkinson, J., Murray, B., Woodhouse, M., Whale, T., Baustian, K., & Carslaw, K., Dobbie, S., O'Sullivan, D., and Malkin, T. L: The importance of feldspar for ice nucleation by mineral dust in mixed-phase clouds, Nature, 498, 355-358, doi:10.1038/nature12278, 2013.

Boucher, O., Randall, D., Artaxo, P., Bretherton, C., Feingold, G., Forster, P., Kerminen, V.-M. , Kondo, Y., Liao, H., Lohmann, U., Rasch, P., Satheesh, S.K., Sherwood, S., Stevens B., and Zhang, X. Y.,: Clouds and Aerosols, Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, 5, 571-657, 2013.

Breiman L.: Bagging Predictors. Machine Learning, 24, 123-140, 1996.

Breiman L.: Random Forests. Machine Learning, 45, 5-32, 2001.

Coe, H., Allan, J. D.: In Analytical Techniques for Atmospheric Measurement; Heard, D. E., Ed., Blackwell Publishing, 265–311, 2006.

Cziczo, D., Thomson, D., Thompson, T., DeMott, P., and Murphy, D.: Particle analysis by laser mass spectrometry (PALMS) studies of ice nuclei and other low number density particles, Int. J. Mass. Spectrom., 258, 21-29, 2006.

Cziczo, D. J., Froyd, K., Hoose, C., Jensen, E., Diao, M., Zondlo, M., Smith, J. B., Twohy, C. H., and Murphy, D. M.: Clarifying the Dominant Sources and Mechanisms of Cirrus Cloud Formation, Science, 340, 1320-1324, doi:10.1126/science.1234145, 2013.

Cziczo, D. J., Thomson, D. S., and Murphy, D. M.: Ablation, flux, and atmospheric implications of meteors inferred from stratospheric aerosol, Science, 291 (5509), 1772–1775, 2001.

Gallavardin, S., Lohmann, U., and Cziczo, D.: Analysis and differentiation of mineral dust by single particle laser mass spectrometry, Int. J. Mass. Spectrom., 274, 56-63, doi:10.1016/j.ijms.2008.04.031, 2008.

Gallavardin, S. J., Froyd, K. D., Lohmann, U., Möhler, O., Murphy, D. M., Cziczo, D. J.: Single Particle Laser Mass Spectrometry Applied to Differential Ice Nucleation Experiments at the AIDA Chamber, Aerosol Sci. Tech., 42, 773-791, doi: 10.1080/02786820802339538, 2008.

Garimella, S., Wolf, M. J., Christopoulos, C. D., Zawadowicz, M. A., and Cziczo, D. J.: Measuring the cloud formation potential of fly ash particle, Atmos. Chem. Phys. (in prep)

Gross, D., Atlas, R., Rzeszotarski, J., Turetsky, E., Christensen, J., Benzaid, S., Olson, J., Smith, T., Steinberg, L., and Sulman, J.: Environmental chemistry through intelligent atmospheric data analysis, Environ. Modell. Softw., 25, 760-769, 2008.

Henning, S., Ziese, M., Kiselev, A., Saathoff, H., Möhler, O., Mentel, T. F., Buchholz, A., Spindler, C., Michaud, V., Monier, M., Sellegri, K. and

Stratmann, F.: Hygroscopic growth and droplet activation of soot particles: uncoated, succinct or sulfuric acid coated, Atmos. Chem. Phys., 12(10), 4525–4537, doi:10.5194/acp-12-4525-2012, 2012.

Hoose, C. and Möhler, O.: Heterogeneous ice nucleation on atmospheric aerosols: a review of results from laboratory experiments, Atmos. Chem. Phys., 12, 9817-9858, doi:10.5194/acpd-12-12531-2012, 2012.

Hiranuma, N., Augustin-Bauditz, S., Bingemer, H., Budke, C., Curtius, J., Danielczok, A., Diehl, K., Dreischmeier, K., Ebert, M., Frank, F., Hoffmann, N., Kandler, K., Kiselev, A., Koop, T., Leisner, T., Möhler, O., Nillius, B., Peckhaus, A., Rose, D., Weinbruch, S., Wex, H., Boose, Y., Demott, P. J., Hader, J. D., Hill, T. C. J., Kanji, Z. A., Kulkarni, G., Levin, E. J. T., McCluskey, C. S., Murakami, M., Murray, B. J., Niedermeier, D., Petters, M. D., O'Sullivan, D., Saito, A., Schill, G. P., Tajiri, T., Tolbert, M. A., Welti, A., Whale, T. F., Wright, T. P. and Yamashita, K.: A comprehensive laboratory study on the immersion freezing behavior of illite NX particles: A comparison of 17 ice nucleation measurement techniques, Atmos. Chem. Phys., 15(5), doi:10.5194/acp-15-2489-2015, 2015a.

Hiranuma, N., Möhler, O., Yamashita, K., Tajiri, T., Saito, A., Kiselev, A., Hoffmann, N., Hoose, C., Jantsch, E., Koop, T. and Murakami, M.: Ice nucleation by cellulose and its potential contribution to ice formation in clouds, Nat. Geosci., 8(4), 273–277, doi:10.1038/ngeo2374, 2015b.

Lesins, G., Chylek, P., & Lohmann, U.: A study of internal and external mixing scenarios and its effect on aerosol optical properties and direct radiative forcing, J. Geophys. Res.-Atmos., 107, 1-12, doi:10.1029/2001jd000973, 2002.

Lohmann, U., and Feichter, J.: Global indirect aerosol effects: a review, Atmos. Chem. Phys., 5, 715-737, doi:10.5194/acp-5-715-2005, 2005.

Lubin, D., and Vogelmann, A.: A climatologically significant aerosol longwave indirect effect in the Arctic. Nature, 439, 453-456, doi:10.1038/nature04449, 2006.

Mjolsness, E.: Machine Learning for Science: State of the Art and Future Prospects, Science, 293, 2051-2055, doi:10.1126/science.293.5537.2051, 2001.

Murphy, D. M.: The design of single particle laser mass spectrometers, Mass Spectrom. Rev., 26 (2), 150–165, 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., 37:4, 382-391, doi:10.1080/02786820300971, 2003.

Niemand, M., Möhler, O., Vogel, B., Vogel, H., Hoose, C., Connolly, P., Klein, H., Bingemer, H., DeMott, P., Skrotzki, J. and Leisner, T.: A Particle-Surface-Area-Based Parameterization of Immersion Freezing on Desert Dust Particles, J. Atmos. Sci., 69, 3077-3092, 2012.

Peckhaus, A., Kiselev, A., Hiron, T., Ebert, M. and Leisner, T.: A comparative study of K-rich and Na/Ca-rich feldspar ice-nucleating particles in a nanoliter droplet freezing assay, Atmos. Chem. Phys., 16(18), 11477–11496, doi:10.5194/acp-16-11477-2016, 2016.

Powers D. W.: Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation, Journal of Machine Learning Technologies, 7, 1-24, 2007.

Saathoff, H., Naumann, K.-H., Schnaiter, M., Schöck, W., Möhler, O., Schurath, U., Weingartner, E., Gysel, M. and Baltensperger, U.: Coating of soot and (NH4)2SO4 particles by ozonolysis products of α-pinene, J. Aerosol Sci., 34(10), 1297–1321, doi:10.1016/S0021-8502(03)00364-1, 2003.

Steinke, I., Funk, R., Busse, J., Iturri, A., Kirchen, S., Leue, M., Möhler, O., Schwartz,T., Schnaiter, M., Sierau, B., Toprak, E., Ullrich, R., Ulrich, A., Hoose, C. and Leisner, T.: Ice nucleation activity of agricultural soil dust aerosols from Mongolia, Argentina, and Germany, J. Geophys. Res. Atmos., doi:10.1002/2016JD025160, 2016.

Vogelmann, A., McFarquhar, G., Ogren, J., Turner, D., Comstock, J., Feingold, G., Long, C., Jonsson, H., Bucholtz, A., Collins, D., Diskin, G., Gerber, H., Lawson, R., Woods, R., Andrews, E., Yang, H., Chiu, J., Hartsock, D., Hubbe, J., Lo, C.,Marshak, A., Monroe, J., McFarlane, S., Schmid, B., Tomlinson, J. and Toto, T.: Racoro Extended-Term Aircraft Observations of Boundary Layer Clouds, Bull. Amer. Meteor. Soc., 93, 861-878, 2012.

Welti, A., Lüönd, F., Stetzer, O., and Lohmann, U.: Influence of particle size on the ice nucleating ability of mineral dusts, Atmos. Chem. Phys., 9, 6929-6955, doi:10.5194/acpd-9-6929-2009, 2009.

Zawadowicz, M. A., Froyd, K. D., Murphy, D. M. and Cziczo, D. J.: Improved identification of primary biological aerosol particles using single particle

mass spectrometry, Atmos. Chem. Phys., doi: 10.5194/acp-2016-1119, 2016.

Table Captions

Aerosol type	FIN Label	Description and/or supplier	Generation method	Sample provided by	Reference
Argentinian	SDAr01	Soil dust collected in La Pampa province, Argentina	Dry-dispersed	KIT	(Steinke et al., 2016)
Chinese	SDMo01	Soil collected from Xilingele steppe, China/Inner Mongolia	Dry-dispersed	KIT	(Steinke et al., 2016)
Ethiopian	VSE01	Soil collected in Lake Shala National Park, Ethiopia (collection coordinates: 7.5 N, 38.7 E)	Dry-dispersed	KIT	N/A
German	SDGe01	Arable soil collected near Karlsruhe, Germany	Dry-dispersed	KIT	(Steinke et al., 2016)
Moroccan	DDM01	Soil collected in a rock desert in Morocco (collection coordinates: 33.2 N, 2.0 W)	Dry-dispersed	KIT	N/A
Paulinenaue	N/A	Arable soil collected in Northern Germany (Brandenburg)	Dry-dispersed	KIT	N/A
ATD	N/A	Arizona Test Dust, Powder Technology, Inc. (Arden Hills, MN)	Dry-dispersed	MIT	N/A
Illite	IS03	Illite NX (Arginotec, Germany)	Dry-dispersed	KIT	(Hiranuma et al., 2015a)
Fly ash	N/A	Four samples of fly ash from U.S. power plants: J. Robert Welsh Power Plant (Mount Pleasant, TX), Joppa Power Station (Joppa, IL), Clifty Creek Power Plant (Madison, IN) and Miami Fort Generating Station (Miami Fort, OH) (Fly Ash Direct, Cincinnati, OH)	Dry-dispersed	MIT	(Garimella, 2016; Zawadowicz et al., 2016)
Na-Feldspar	FS05	Sodium and calcium-rich feldspar, samples provided by Institute of Applied Geosciences, Technical University of Darmstadt (Germany) and University of Leeds (UK)	Dry-dispersed	KIT	(Peckhaus et al., 2016)
K-Feldspar	FS01	Potassium-rich feldspar, samples provided by Institute of Applied Geosciences, Technical University of Darmstadt (Germany) and University of Leeds (UK)	Dry-dispersed	KIT	(Peckhaus et al., 2016)

Agar	**N/A**	Agar growth medium for bacteria, Pseudomonas Agar Base (CM0559, Oxoid Microbiology Products, Hampshire, UK)	Wet-generated	KIT	N/A
Bacteria	PS32B74 + PFCGina01	Two different cultures of *Pseudomonas syringae*.	Cultures grown on the agar growth medium (as above), suspended in nanopure water and wet-generated	KIT	(Zawadowicz et al., 2016)
Cellulose	MCC01, FC01	Microcrystalline and fibrous cellulose (Sigma Aldrich, St. Louis, MO)	Wet-generated	KIT	(Hiranuma et al., 2015b)
Hazelnut	PWW-hazelnut	Natural hazelnut pollen (GREER, Lenoir, NC) wash water	Wet-generated	KIT	(Zawadowicz et al., 2016)
Snomax	Snomax	Snomax, (Snomax International, Denver, CO) irradiated, desiccated and ground *Pseudomonas syringae*	Wet-generated	KIT	(Zawadowicz et al., 2016)
PSL	N/A	Polystyrene latex spheres (Polysciences, Inc. Warrington, PA), various sizes	Wet-generated	MIT	N/A
Soot	CAST minOC or maxOC	CAST soot	miniCAST flame soot generator (manufactured by Jing Ltd Zollikofen, Switzerland)	KIT	(Henning et al., 2012)
SOA	SOA	Secondary organic aerosol	Ozonolysis of α-pinene	KIT	(Saathoff et al., 2003)
K-Feldspar cSA	FS01cSA or FS04cSA	Potassium-rich feldspar (as above) coated with sulfuric acid (SA).	Small amounts of sulfuric acid were incrementally added to the chamber filled with K-feldspar to achieve thin coatings, as judged from PALMS spectra	KIT	(Saathoff et al., 2003)
K-Feldspar cSOA	FS04cSOA	Potassium-rich feldspar (as above) coated with secondary organic aerosol (SOA, as above).	Small amounts of SOA were incrementally added to the chamber filled with K-feldspar to achieve thin coatings, as judged from PALMS spectra	KIT	(Saathoff et al., 2003)

Table 1. Description of aerosol types used in training data set. ~~Rows are grouped and colored by broad aerosol categories in the following order: Fertile Soil, Mineral/Metallic, Biological, and Other.~~

Aerosol Type				Broad Categories			
Negative		Positive		Negative		Positive	
ion	feature	ion	feature	ion	feature	ion	feature
35	$^{35}Cl^-$	23	Na^+	35	$^{35}Cl^-$	23	Na^+
25	C_2H^-	59	$Co^{+(1)}/CaF^+/C_2H_2OOH^+$	26	$CN^-/C_2H_2^-$	59	$Co^{+(1)}/CaF^+/C_2H_2OOH^+$
24	C_2^-	39	$^{39}K^+$	46	NO_2^-	44	$SiO^+/COO^+/^{44}Ca^+/AlOH^+$
57	C_2OOH^-	12	C^+	1	H^-	39	$^{39}K^+$
59	$C_2H_2OOH^-/AlO_2^-$	24	C_2^+	57	C_2OOH^-	28	Si^+/CO^+
43	HCN^-/AlO^-	41	$^{41}K^+/C_3H_5^+$	59	$C_2H_2OOH^-/AlO_2^-$	41	$^{41}K^+/C_3H_5^+$
1	H^-	204-208	Pb region (^{204}Pb, ^{206}Pb, ^{207}Pb and ^{208}Pb)	45	$COOH^-$	54	$^{54}Fe^+$
26	$CN^-/C_2H_2^-$	27	$Al^+/C_2H_3^+$	42	$CNO^-/C_2H_2O^-$	56	Fe^+/CaO^+
46	NO_2^-	44	$SiO^+/COO^+/^{44}Ca^+/AlOH^+$	43	HCN^-/AlO^-	27	$Al^+/C_2H_3^+$
16	O^-	57	$^{57}Fe^+/CaOH^+/C_3H_4OH^+$	16	O^-	45	$SiOH^+/COOH^+$
17	OH^-	N/A	aerodynamic diameter	73	$C_2O_3H^-/C_3H_2OOH_3^-$	66	Zn^+
61	$SiO_2H^-/^{29}SiO_2^-/C_5H^-/CHO_3^-$	83	$H_3SO_3^+/C_4H_2OOH^+$	63	PO_2^-	57	$^{57}Fe^+/CaOH^+/C_3H_4OH^+$
63	PO_2^-	87	$^{87}Rb^+/CaPO^+$	60	$SiO_2^-/C_5^-/CO_3^-/AlO_2H^-$	87	$^{87}Rb^+/CaPO^+$
19	F^-/H_3O^-	13	CH^+	15	NH^-/CH_3^-	85	$^{85}Rb^+$
76	SiO_3^-	66	Zn^+	24	C_2^-	83	$H_3SO_3^+/C_4H_2OOH^+$
77	$SiO_3H^-/^{29}SiO_3^-$	28	Si^+/CO^+	76	SiO_3^-	24	C_2^+
79	PO_3^-	85	$^{85}Rb^+$	32	O_2^-	204-208	Pb region (^{204}Pb, ^{206}Pb, ^{207}Pb and ^{208}Pb)
60	$SiO_2^-/C_5^-/CO_3^-/AlO_2H^-$	72	FeO^+/CaO_2^+	N/A	aerodynamic diameter	40	Ca^+
45	$COOH^-$	54	$^{54}Fe^+$	71	$C_3H_2OOH^-$	153	$^{137}BaO^+$
N/A	aerodynamic diameter	82	ZnO^+	50	$C_4H_2^-$	N/A	aerodynamic diameter

[(1)] Contamination

Table 2. Features rankings for differentiation of particles between labels and between broad categories in positive and negative ion modes. See text for additional details.

Category	Negative	Postive
Fertile Soil	0.88	0.83
Mineral/Metallic	0.93	0.98
Biological	1.00	1.00
Other	0.96	0.93

Category	Negative	Postive
Fertile Soil	0.024 ±0.020	0.035 ±0.033
Mineral/Metallic	0.017 ± 0.027	0.006 ±0.008
Biological	0.000	0.001 ± 0.002
Other	0.021 ±0.015	0.024 ±0.053

Table 3. Model performance by category and ion mode on a population consisting entirely of aerosols within that category. Left: Average classification accuracy where 1.0 = 100% precision (Powers, 2007). Right: mean and standard deviations of misclassification.

Figure

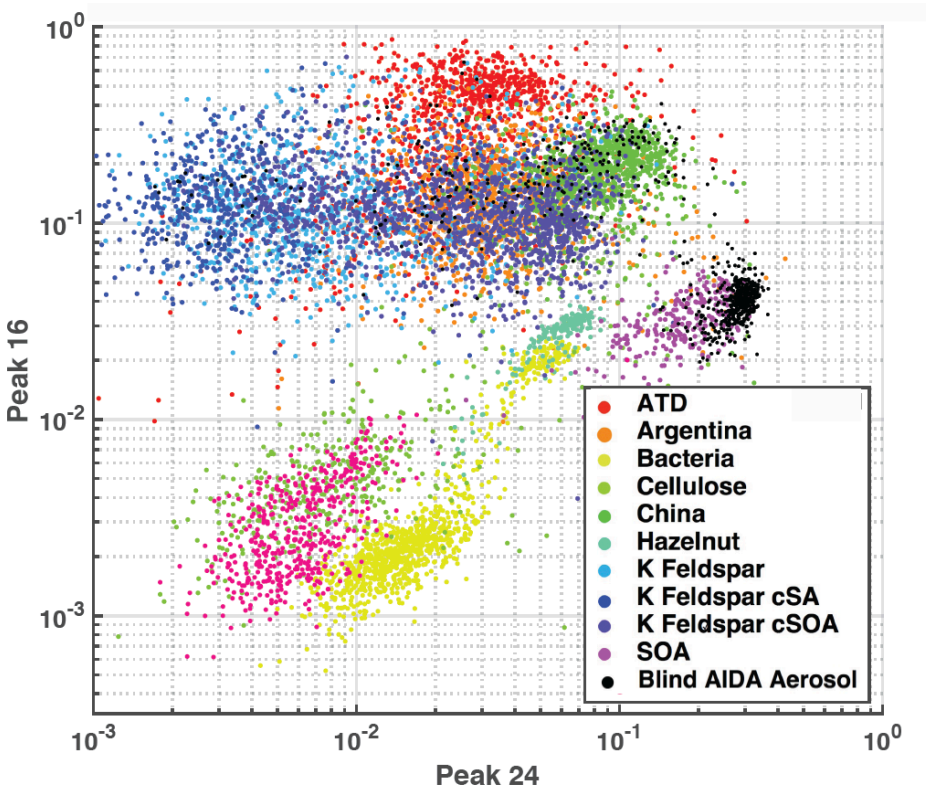


Figure 1: Aerosol training data plotted as feature area 16 (O^-) verses area 24 (C_2^-). Axes represent peak areas normalized to total signal obtained from PALMS (i.e., 1 = 100% of signal). This illustrates simple 2-dimensional clustering of aerosols from the training data set by type. Co-plotted are ~500 randomly drawn spectra from the AIDA blind experiment, which were known to be a subset of the training data aerosols.

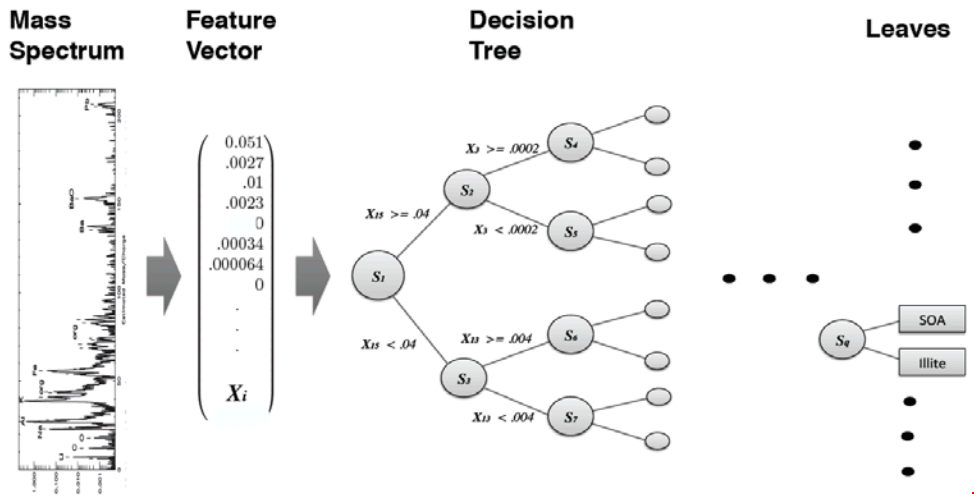
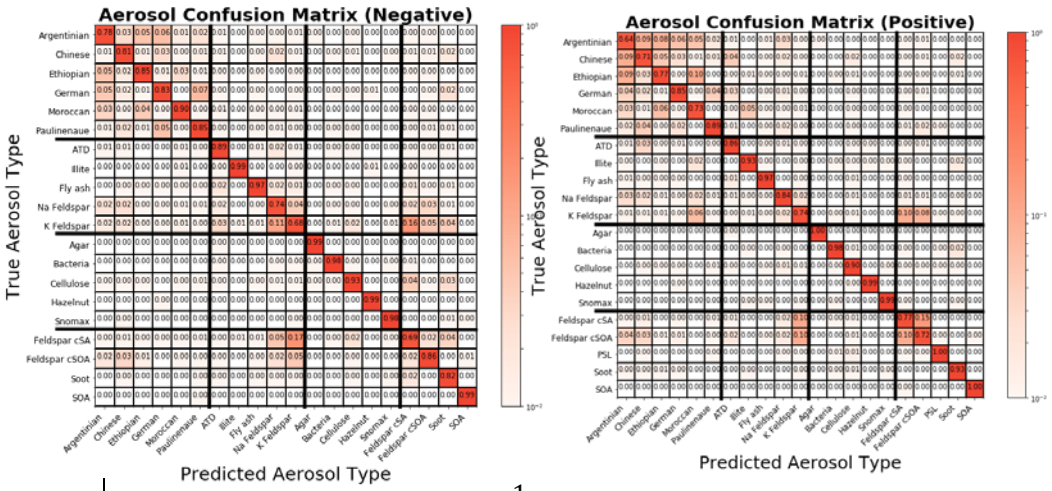


Figure 2. Schematic of decision tree classification for a single aerosol spectrum. From left to right, a mass spectrum is normalized with respect to total ion current, forming the elements of normalized feature vector X. A trained decision tree then applies a series of tests to a discreet number of peaks in order to arrive at a categorical aerosol prediction (the leaves).



1

2

3 Figure 3. Column-normalized confusion matrices showing fraction of aerosols labeled as
4 j that belong to i, where i and j are row and column indices, respectively. Confusion
5 matrices are determined from training data of known origin and are used to compute
6 probability distributions. Aerosol types (Table 1.) are grouped into four broad categories
7 delineated by the bold horizontal and vertical bars. From top to bottom or left to right:
8 fertile soils, mineral/metallic, biological, and other. Classification accuracy, the average
9 probability of a correct aerosol prediction across all labels, is computed by averaging
10 diagonal matrix elements. For all aerosol types, the accuracy is ~~87~~88% in positive ion
11 mode and ~~87~~86% in negative ion mode.

12

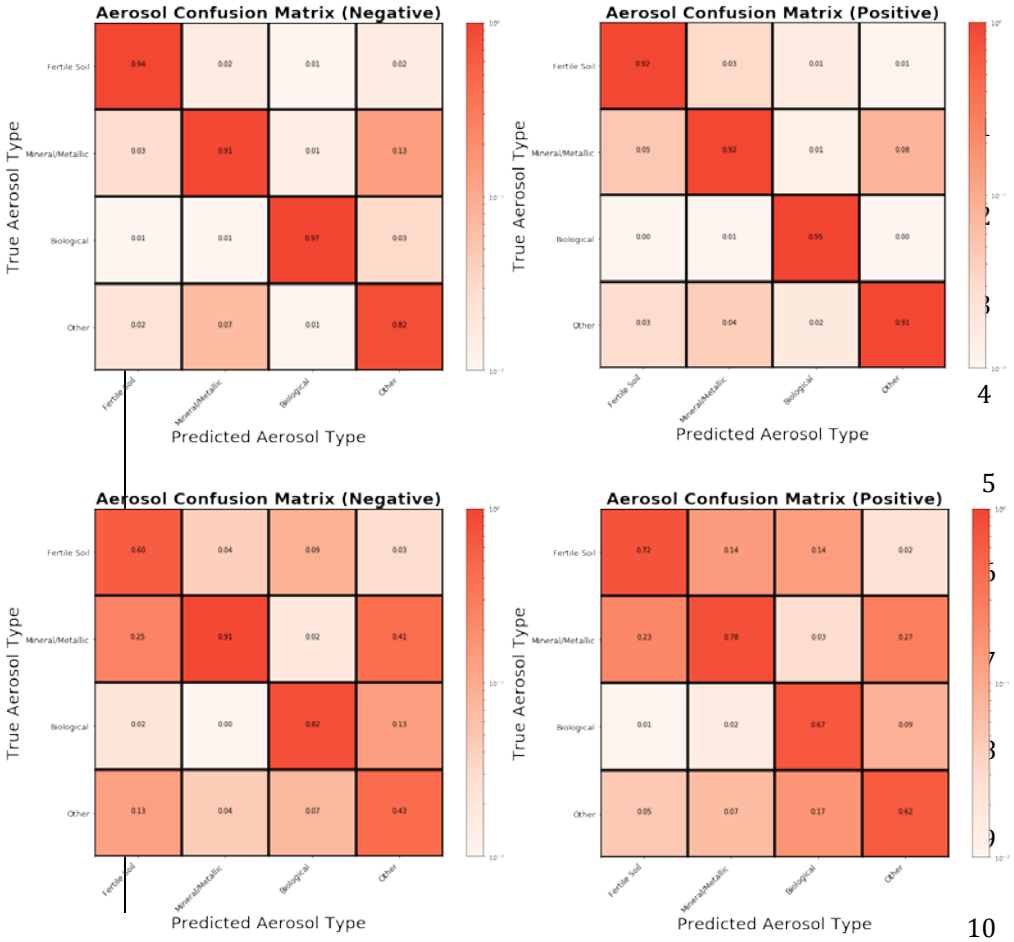


Figure 4. Column-normalized confusion matrices for the broad categorization of aerosols following the convention in Figure 3. ~~a.)Top row:~~ For all aerosol categories, the ~~random forest has an~~ accuracy ~~of 93~~is 94% in positive ion mode and ~~91~~ 92% in negative ion mode. ~~b.) Bottom row: The Euclidean distance classifier has an accuracy of 70% in positive ion mode and 69% in negative ion mode~~

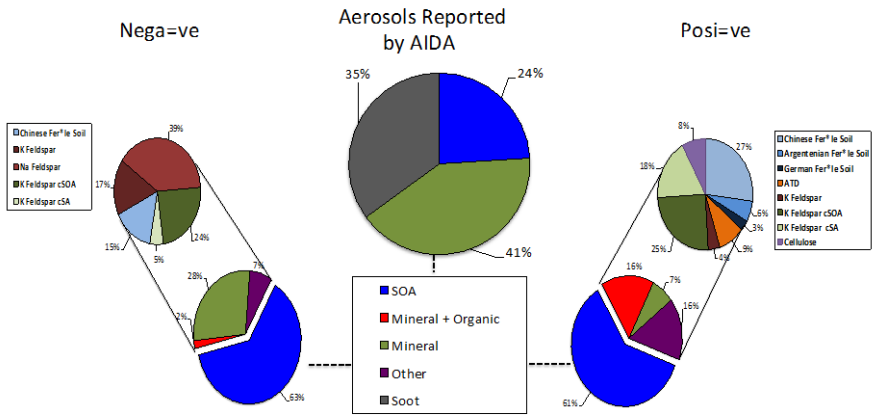


Figure 5. Model predictions of ~5000 aerosols sampled from the AIDA FIN01 blind mixture which was known to be a subset of the training data. Top middle : aerosol types input to the chamber for the blind mixture. Model predictions are shown for negative and positive ion mode on the left and right, respectively. Bottom: broad categories. Top: breakout by aerosol type of the non-SOA categories above the 1% level. Notes (1) the soot in the blind mixture was known to be below the instrument detection limit and therefore is not expected to be found in the data, (2) coagulation of SOA and mineral dust, which occurred after aerosol input to the chamber, ~~was often categorized as mixed mineral and organic particles or fertile soils (i.e., mixtures of mineral and organic components) considered in the training data set, (3) the aerosols types reported by AIDA do not account for PALMS transmission efficiency (see text for details).~~ appears as the mineral + organic category.