Atmos. Chem. Phys. Discuss.,
https://doi.org/10.5194/acp-2020-868-RC1, 2020

Atmospheric Chemistry and Physics Discussions

Interactive comment on "Eight years of sub-micrometre organic aerosol composition data from the boreal forest characterized using a machine-learning approach" *by* Liine Heikkinen et al.

Anonymous Referee #1

Received and published: 24 September 2020

This paper presents a factorisation analysis of a very long time series of ACSM data from the Hyytialla site. While PMF analysis has worked well for short AMS and ACSM datasets, multi-season and multi-annual datasets have previously necessitated breaking the dataset down into small chunks or using a 'rolling window' method. The approach here is to use a 'rolling relaxed CMB' approach, with the a priori mass spectral profiles generated using the unsupervised clustering of bootstrapped 'rolling PMF'. This is relevant for ACP, because while this paper does have a heavily technical bent, it also reports general interpretations of the driving factors behind aerosol behaviour at this

heavily-studied site that may have implications for other studies.

While this paper presents a very detailed account of what was done in terms of analysis, I do not find myself completely convinced that the authors have sufficiently argued the case for why this approach should be considered superior to other existing techniques. To be clear, I am not necessarily saying that they should have done anything differently (I would particularly commend the adherence to unsupervised methods for the sake of objectivity), but my opinion is that the authors need present stronger arguments for why certain choices were made in the approach to data analysis and they also need to go further in exploring the strengths and weaknesses of this approach (compared to others) rather than simply accept the outputs at face value. It's also not completely clear to me how different the interpretation of the results has been positively influenced by the use of the new technique. Put simply, it's not very clearly spelled out what the authors were trying to achieve with this approach, what the underlying principles were (as opposed to the detail of the algorithms), and whether they objectively succeeded in meeting their original objectives. For these reasons, I am recommending publication subject to major revisions. I would ask the authors to consider the points below.

Major comments:

It took me several reads before I really thought I understood the philosophy behind this technique, as there are a frankly baffling number of analysis stages. The authors need to be much clearer in how they describe the approach and I don't think presenting it in a stepwise form in figure 1 really helps. Instead, it should be made clear from the beginning that (assuming I did understand it correctly) the ultimate goal is to generate factorisation outputs using rolling rCMB and the biggest emphasis should be placed on this technique. As far as I can tell, all of the other stages leading up to this are merely generating objective mass spectral profiles that this algorithm uses as an input.

The use of clustering represents the part of the analysis that is traditionally used least in ACSM analysis and therefore represents the part with the least precedent. Certain decisions are made at various stages of the clustering and while the authors present arguments for why these would be considered reasonable, they do not really explore the notion of what would happen if they had done it in a different manner. I would specifically point to: 1. The use of k-means as a clustering algorithm. 2. The use of cosine angle rather than Euclidian distance as the distance measure. 3. The weighting of higher mass to charge ratios. 4. The choice of methods used to determine the optimum number of clusters and centroid profiles. While I am not questioning the individual choices, I would expect the authors to offer more reasons why alternatives were discounted, reporting on any undesirable behaviour when alternatives were test, where available.

I would consider the benchmark comparison of this technique to be rolling PMF, as this is an already-developed technique that is used to interpret long ACSM datasets. In figure 4, the authors compare weighted residuals and the reader could be forgiven for thinking that this has resulted in an inferior data product to what was obtained at the end of stage 1 of the analysis. The authors need to discuss in more depth the pros and cons of each approach and present a stronger case for why it should be considered advantageous to use the technique used here. This should bear in mind that just because a particular algorithm uses less supervision and produces a less ambiguous result, these do not in themselves mean the data products are intrinsically more accurate.

As part of any numerical data reduction such as this, it is vital to properly explore its limitations and I don't consider what is presented here to be sufficient. I am particularly interested in the analysis of residuals, which is normally the first thing to inspect. Is the unexplained variance and mass shown in figure 8 purely random noise, or is there any structure (relative to m/z, season, temperature, time of day, etc.) that might suggest there are factors at work that this does not adequately capture?

A key detail in the final data products is by how much the mass spectral profiles of the

factors was allowed to vary in the rCMB analysis. As I understand it, if the profile could vary with the rolling window, then this would avoid many of the limitations imposed by the default PMF data model, however this then creates implications for how the results are interpreted, in particular with the seasonal analysis and long term trends. Can the authors be sure that any interpretations presented in section 6.2 are the result in changes in the abundance of the different organic aerosol types, or changes in the mass spectral profile? If, on the other hand, the profiles are rigid, then this instantly asks questions of whether the factorisation is equally applicable at all times, or whether the technique is still susceptible to the same rotational ambiguity problems as conventional PMF.

Generally speaking, the explanations of the behaviour in section 6.2 are largely speculative and seem to be focused on creating a plausible narrative rather than offering new scientific insight. The authors should focus the discussion on what this new work adds to the (already substantial) body of work concerning this site and tropospheric aerosol processes in general, specifically through the virtue of adopting this new technique (as opposed to other existing methods). While hypotheses are frequently referred to casually in the text, it's not completely clear how these are being tested by the data and to what certainty.

Minor comments:

The term 'openair' is used to identify certain types of plots, but openair represents a large suite of many different graphing tools. Furthermore, openair wasn't actually used to generate the ones here. While it may be appropriate to credit the development of openair with popularising these graph types, the plots should be referred to by their specific type, e.g. polar plots.

I found section 6 overly wordy, with more text than was necessary to convey the important information. The authors may wish to cut back on the amount of discussion presented concerning the approach to analysis, instead focusing on conveying the new scientific insights this offers, ideally in the form of hypothesis testing.

One of the problems that has traditionally confounded long term PMF analysis is that the mass spectral profiles of OOA factors can vary with season. Did the authors find any evidence of this at any stage of the analysis?

Regarding figure 6, the overlaid ellipses are not necessary, given that the points are already coloured. I would remove them, as they only serve to distract.

Also regarding figure 6, a number of points associated with LV-OOA have a high f60, which is a classic symptom of LV-OOA ‘mixing’ with BBOA, owing to the high HULIS content of the latter. The authors should comment on this.

Section 6.2.4: A statistical treatment is referred to, but no actual quantitative results are presented. This should be done, even if it is to report that no significant trend was observed.

Interactive comment on Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2020-868, 2020.