Atmos. Chem. Phys. Discuss.,
https://doi.org/10.5194/acp-2020-868-RC2, 2021


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 #2**

Received and published: 18 January 2021

Atmospheric Chemistry and Physics Manuscript ID: acp-2020-868 Title: Eight years of sub-micrometre organic aerosol composition data from the boreal forest characterized using a machine-learning approach

General comments:

This manuscript describes an interesting approach to deal with the source apportionment of long submicron organic aerosol (OA) datasets at a remote site (SMEAR II research supersite in the boreal forest). At these types of sites, far from primary sources,



a "classical" rolling PMF method (commonly based on the identification and constraining of primary factors through a criteria-based approach) may not be sufficient to deal with mostly only secondary factors.

This article presents a new methodological approach based on K-means clustering of a very large number of unconstrained rolling PMF solutions, to overcome this difficulty. The extracted clusters are further used to run a CMB-type (constrained) rolling PMF. Output results are finally utilized to provide an overview of the temporal variability of the three OA factors identified. Methodological results (section 5) extend from pages 11 to 21 (with about 3 pages of Figures), while the "Results" section (section 6) ranges from page 21 to 37 (with about 7 pages of Figures). The general impression is that the paper permanently hesitates (and authors have been unable to choose) between the two focal points (methodology- or result-oriented) which is confusing and frustrating at times, especially when the methodological choices seem too descriptive (see specific comments below).

The overall quality of the manuscript is excellent. A few minor technical corrections are reported at the end of this document.

**********

Specific comments:

L40-41: "it is also possible that ACSM was less efficiently capturing short term (POA) pollution plumes". I believe this assumption derives from the behavior observed with the 3-hour averaged dataset. Therefore could it rather arise from the averaging (diluting the information over a longer time window) compared to the ACSM raw time resolution (and even more if compared to high time resolution AMS measurements performed at the same site) rather than from sampling differences like "capturing" may suggest?

L171 & L174: "Open air (. . .) analyses" and "Openair polar plots" are not informative titles and should be modified. Maybe something like "Influence of meteorological pa-

rameters on air pollutant concentrations" for section 3, and "OA variation with wind direction and speed" for subsection 3.1 (or something along that line)? See also section 6.3.1 (L789).

L183: In the backtrajectory and TOL analysis, were the possible rainfalls along the air mass trajectory taken into account since it could highly influence the aerosol loading arriving at the site?

L323: Considering this is a new method, this section should not only be descriptive, but also provide some elements to compare between unconstrained rolling PMF/constrained rolling PMF/machine-learning approach ... and highlight the advantages of this latter methodology – at least in this case – to provide a more accurate analysis of the OA fraction. It is not so obvious to me after reading the manuscript what logic was behind the different validation steps (which are not so explicit).

L330: Since a correct assessment of the error matrix is critical for PMF analysis, please describe (at least in the SI) how the error propagation was performed for the 3-hour averaged OA.

L332: The most common approach uses a step function to downweight the weak and bad variables, based on averaged values for each m/Q. Although I agree with the authors that the cell-wise function may be more appropriate, especially if there is some strong seasonal variability for specific m/Q, have you tested both types of weighing to estimate the difference?

L339: If I understand correctly, since the initial rolling PMF is unconstrained, it is possible to have three factors that vary over the seasons/years. For instance, BBOA, HOA and 1 OOA in winter then switching gradually to HOA and 2 OOAs. But what if only two factors are relevant for some runs? Did you have a way to account for that (and eventually discard those runs)? It is relatively common in PMF analyses to keep only the "best" runs and discard those who do not fit criteria. Here I am not sure to understand if that happens in that first step (I believe not) or if using the weighted cluster centroids



in the second phase is indeed taking care of the possible outliers.

L359 (Figure 1): I liked the summary of all the steps in that Figure but it would be interesting to make the link with the corresponding sections in the text, either directly on the Figure or at least in the caption.

L394: There could also be strong arguments against mass-to-charge scaling for this type of measurements. Because of their low signal-to-noise ratios, high m/Qs are often downweighted (as weak variables) in the PMF analysis, even if they can contain more information on OA sources. Have you selected only those for which the SNR was high enough in the dataset? Otherwise it seems contradictory to the previous reasoning applied to PMF. And I am wondering if it could play a role in the slight overestimation of POA in summertime when applying the rCMB.

L521 and following, and Fig. S.4: What is the proportion of "the PMF windows where POA was not classified"? The number of runs for which this factor appears compared to the two others could be indicated in Fig. S.4 as well (I guess it is N = 28, mentioned in Fig. 5e which appears later in the text?). What is puzzling me is that POA should appear more "easily" in the rCMB PMF runs when its concentration is higher, and clearly this is what will drive the slope of the linear fit in Fig. S.4c. Here the slope between the two POA factors (from clustering and rCMB) stays low (0.48) despite quite a few points at high concentrations.

L650: Stronger conclusions on the sources (anthropogenic or not; local or not) of LV-OOA could be drawn from various wind data analyses (backtrajectories, NWR plots, etc.). Thus it feels strange to find them here since section 6.3 specifically deals with the wind sectors influencing the sampling site. I would suggest to shorten the discussion here and refer to section 6.3 where you should be able to give stronger conclusions.

L713: "a larger wintertime fPOA:". Confusing. I thought the discussion had switched to POA absolute concentrations and not fPOA anymore?

L725: I am not convinced by the usefulness of this section, which I feel unconnected to the rest of the story. The authors themselves state that the dataset is probably not long enough to conduct a robust trend analysis, and that their assessment is likely biased by missing data for some seasons, or technical changes in the sampling line (no dryer for the two earlier years). Besides, this is clearly indicated in the conclusion as well (L916-918).

**********

Technical corrections:

L33: “However, also the nearby”. Please consider revising this sentence.

L114: “recorded”

L119: “a new framework”

L166: Please provide the CPC model.

L238-240: Unclear. Please consider revising this sentence.

L297: “Reassigning all the points”

L317: “indicate a good”

L322-323: “a relaxed (. . .) PMF analysis” (or another singular term)

L347: “28 days”. I believe it is not hours.

L347: “Only window widths”

L367 (Figure 2): please specify “time series of 3-hour averaged OA”

L509: “such a dynamic”

L512: remove “rolling”

L591-592: “we can see that the our SV-OOA O:C”. Delete “the”.



L627-628: “in in”. Delete one.

L663 and other occurrences in the text: “diel cycle” would be a more appropriate term than “diurnal cycle”.

L671-672: seems like a sentence fragment. Consider revising.

L689: “is does likely not play”. Please revise.

L748: “A-OA”. Do you mean “POA”?

L757 (Figure 7): I honestly do not see a great difference between the light and dark shadings. Could you add more contrast?

L764: “POA” in the caption does not seem vertically aligned with the others.

L863: “As these figures, do not indicate”. Delete the comma.

L869 (Figure 12a): IQR overlap in this plot. May be more readable with one plot per factor.

L895: “that the need (. . .) to increase”. Please revise.

L1225-1230: Two references not in alphabetical order.

Supplementary information

L33: “The coloured lines represent”

Figure S6: Plots would be more readable if vertically expanded. Coloured and grey shaded areas are not explicit in the caption.

Figure S7: IQR overlap in these plots. They may be more readable with one plot per factor.

---

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