Supplementary information

Eight years of organic aerosol composition data from the boreal forest characterized using a machine-learning approach

Liine Heikkinen[1], Mikko Äijälä[1], Kaspar Daellenbach[1], Gang Chen[2], Olga Garmash[1], Diego Aliaga[1], Frans Graeffe[1], Meri Räty[1], Krista Luoma[1], Pasi Aalto[1], Markku Kulmala[1], Tuukka Petäjä[1], Douglas Worsnop[1,3], and Mikael Ehn[1]

[1]Institute for Atmospheric and Earth System Research /Physics, Faculty of Science, University of Helsinki, Helsinki, FI–00014, Finland
[2]Laboratory of Atmospheric Chemistry, Paul Scherrer Institute, Villigen, Switzerland
[3]Aerodyne Research Inc., Billerica, MA, USA

Correspondence to: Liine Heikkinen (liine.heikkinen@helsinki.fi) and Mikael Ehn (mikael.ehn@helsinki.fi)

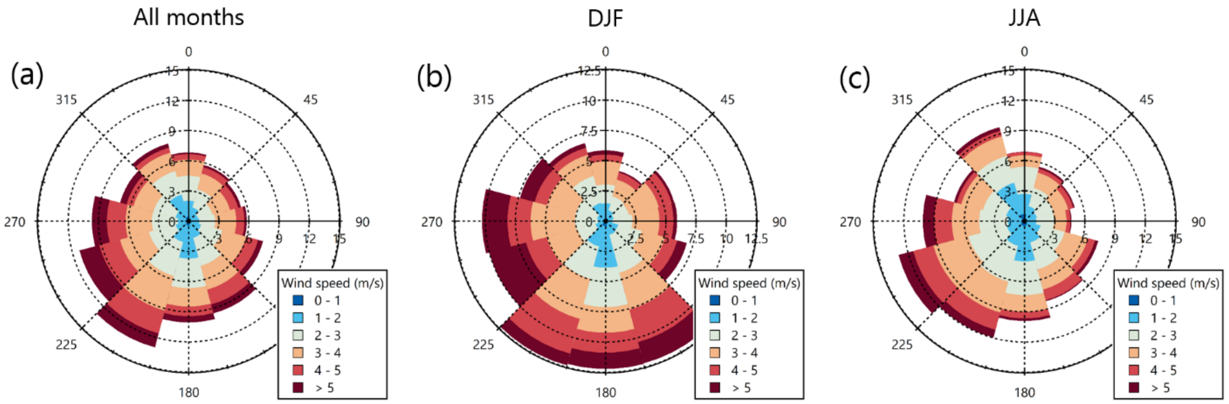


Figure S.1 Note that b-panel probability isolines only reach 12.5% whereas in panels a and c the circle radius is 15%. The wind direction and speed data are collected above the boreal forest canopy.

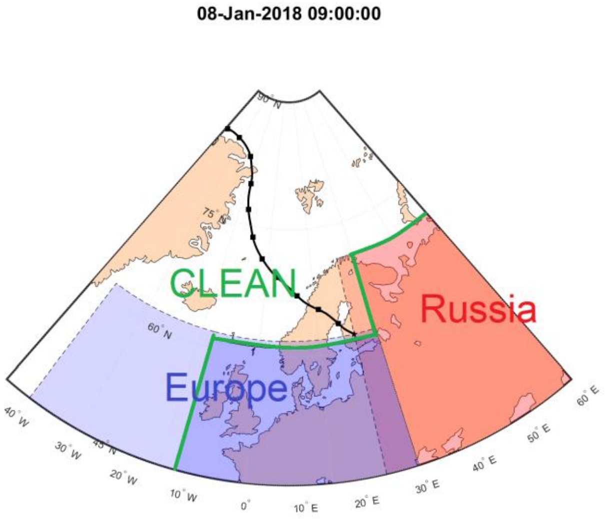


Figure S.2 Sector classification used in the time over land analysis. The red Russia-sector and blue Europe-sector are considered as polluted sectors while the clean sector has the least anthropogenic influence.

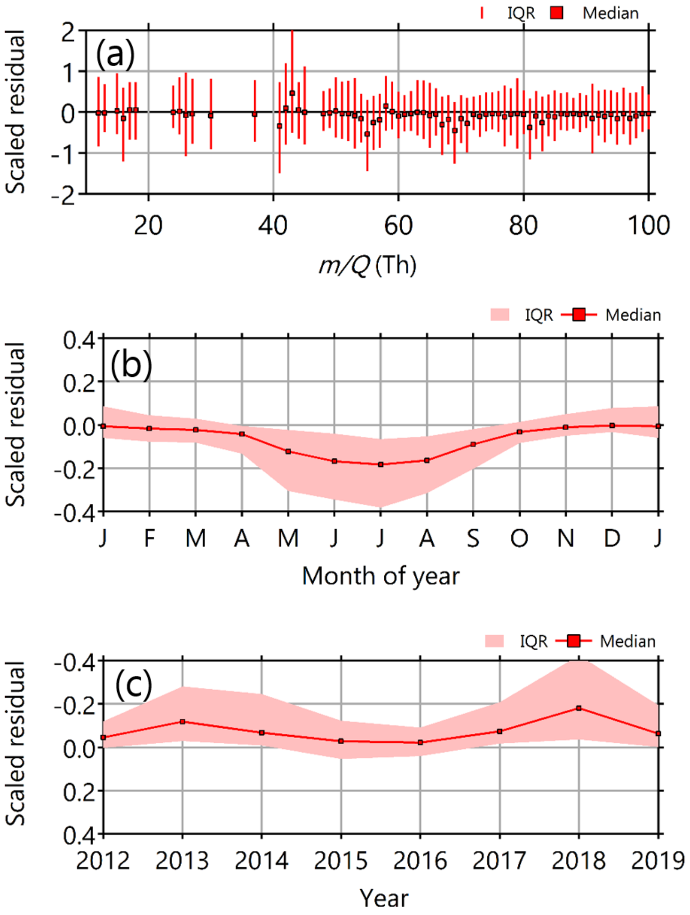


Figure S.3 The residual mass spectrum (panel a), annual cycle of the scaled residual (panel b) and year-to-year variability of the scaled residual (panel c). The a-panel interquartile range (IQR) has large variation making it challenging to recognize patterns in the residual mass spectrum. The monthly cycle of the scaled residual suggests little rCMB overestimation in summer months, which could be linked to the POA overfitting speculated in the main text to occur in summer. The year-to-year variability shows no systematic trends. Years 2013 and 2018 have slightly higher scaled residuals than other years (rCMB underestimation).

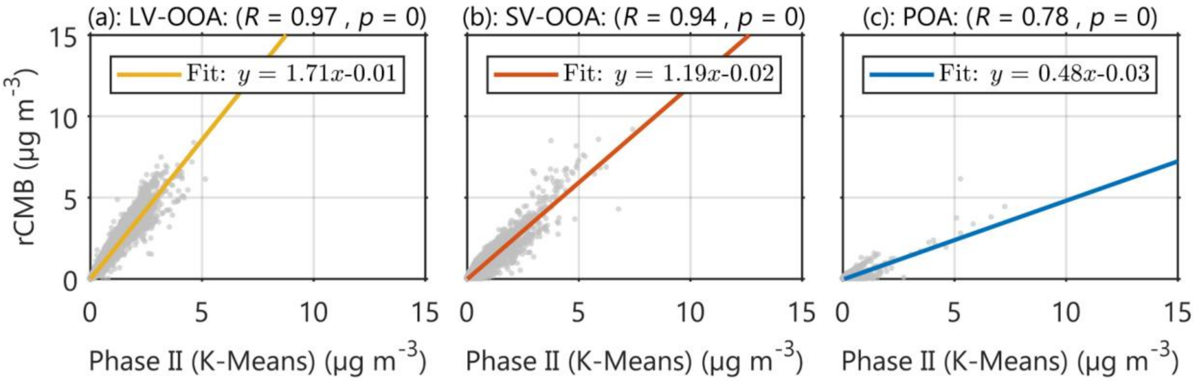


Figure S.4 Comparison of time series obtained after Phase II (K-Means) and rCMB for LV-OOA (panel a), SV-OOA (panel b) and POA (panel c). The coloured lines represents linear fits, and their equations are written in the panel captions. The panel titles contain Pearson correlation coefficient values (*R*) for each plot. The high correlation between the time series supports the decision in obtaining the Phase II (K-Means) cluster centroid temporal behaviour via rCMB.

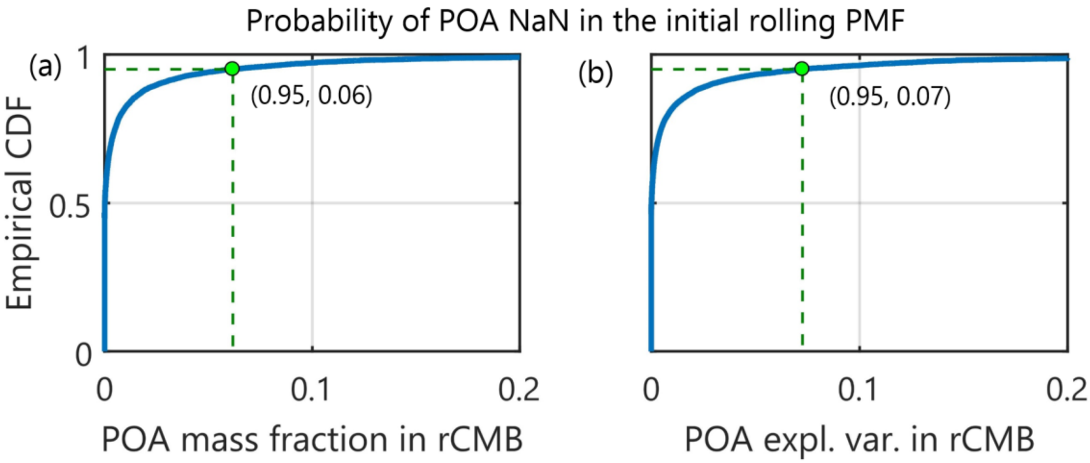
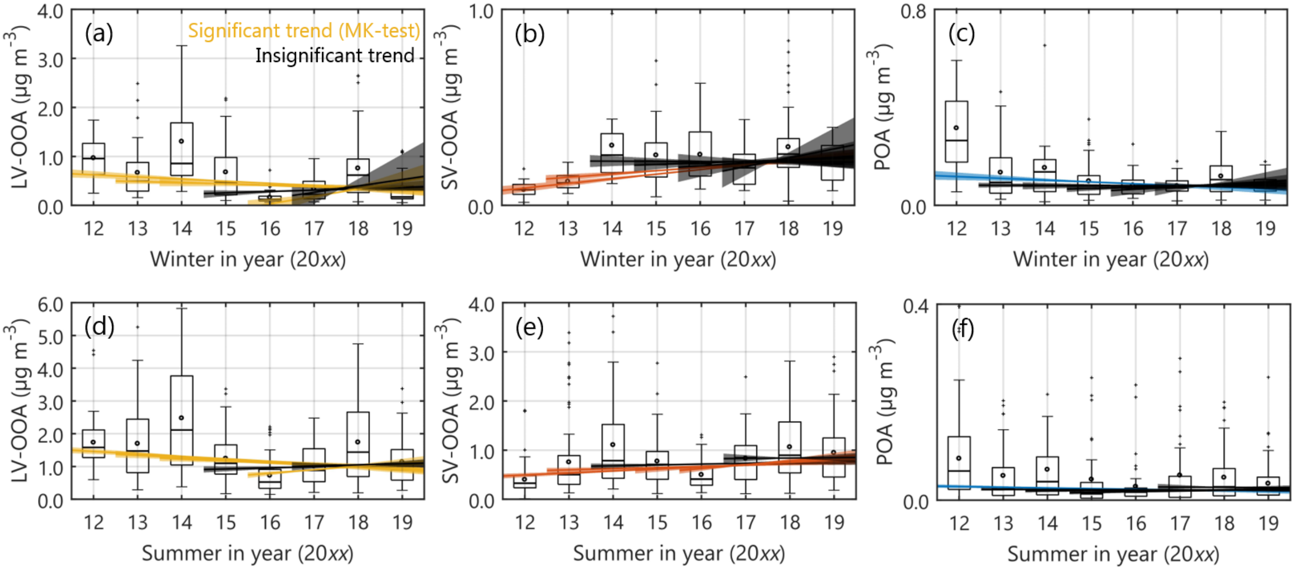


Figure S.5 The cumulative distribution functions showing the POA mass fraction calculated with rCMB when POA was not resolved in the unconstrained initial rolling PMF (i.e. when POA in the rolling window was NaN). The 95th percentile (3σ) is presented as the green marker. It corresponds to a POA mass fraction of ca. 6%. The b-panel holds POA explained variation in the x-axis. The 3σ explained variation is 7%.



FigureS.6 Box plots visualizing median statistics of wintertime (upper row) and summertime (lower row) OA type concentrations. The mean concentrations are drawn with black open circles. The first column represents LV-OOA, second SV-OOA and third POA. The line fit represents the Sen's linear fit. The fits considered as significant with MK-test are drawn in coloured lines while the insignificant trends with black lines. Note different scales in *y*-axes. The grid lines are drawn every 1 µg m^{-3}.

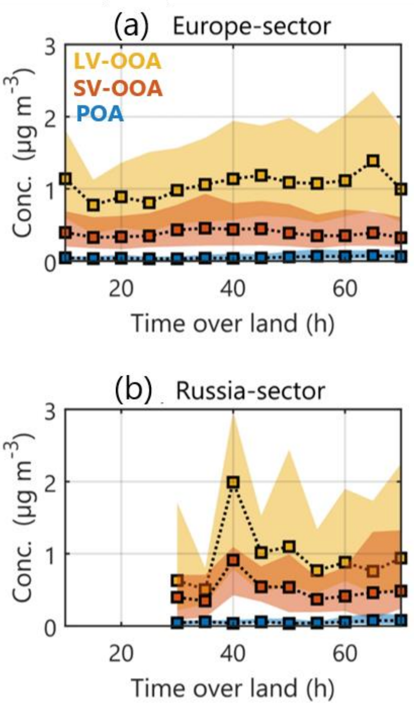


Figure S.7 The different rCMB factors (*y* axes in µg m^{-3}) vs TOL (*x* axes in hours) for the Europe-sector (panel a) and Russia-sector (panel b; see Fig. S.2 for a more precise sector definition). The data are binned to 5-hourly TOL bins. The shaded areas represents the concentration interquartile ranges (25th to 75th percentile) and the square markers the median concentrations.

55