Supplement of Atmos. Chem. Phys., 21, 10081–10109, 2021
https://doi.org/10.5194/acp-21-10081-2021-supplement

Supplement of

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

Liine Heikkinen et al.

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

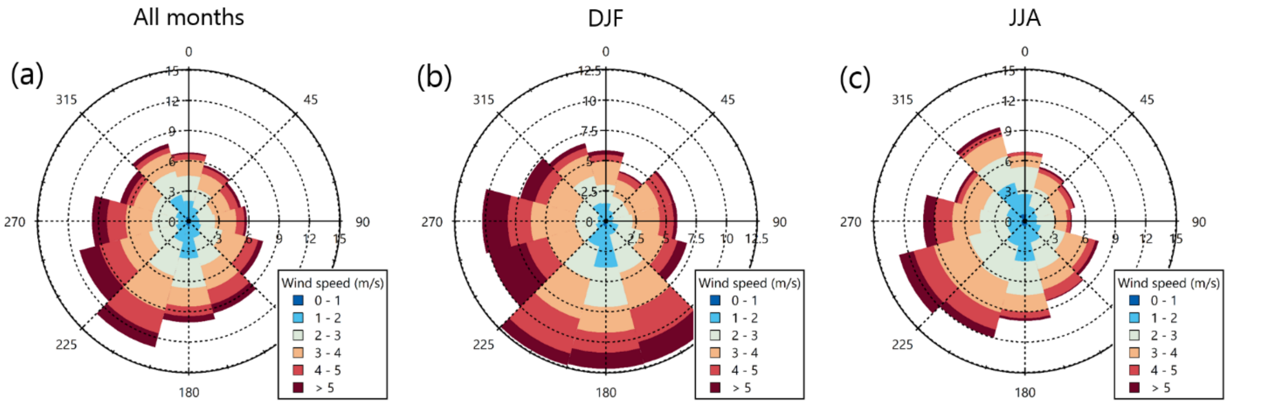

Figure S.1 Note that b-panel probability isolines only reach 12.5% whereas in panels a and c the circle radius is 15%. The wind direction and speed data are collected above the boreal forest canopy.

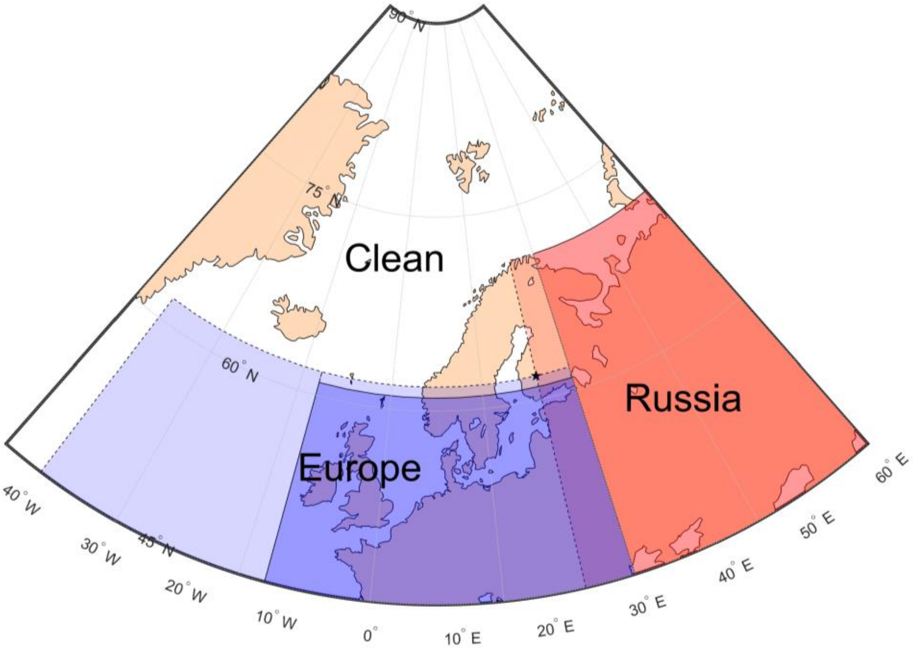

Figure S.2 Sector classification used in the time over land analysis. The location of SMEAR II station is indicated with a star symbol. The red Russia-sector and blue Europe-sector are considered polluted, while the clean sector has the least anthropogenic influence. The sectors overlap in the areas with higher color transparency that are outlined by dashed lines.

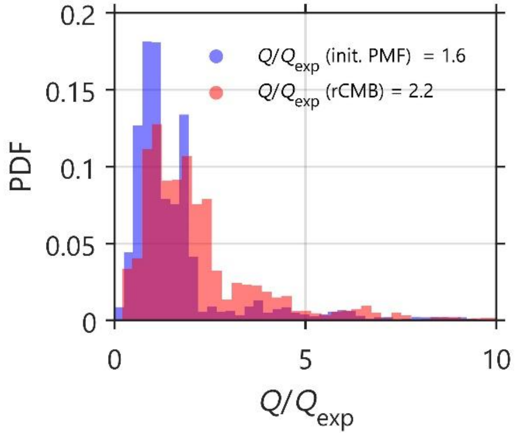

Figure S.3 Median silhouette scores calculated for rCMB scaled residual matrix *k*-means clustering attempts for 2–9 clusters. The residual matrix was mass scaled similarly to the PMF output pass spectral matrix prior to the clustering. No structures could be found within the data which is revealed by the low median silhouettes as well by the fact that convergence within the *k*-means was often not achieved.

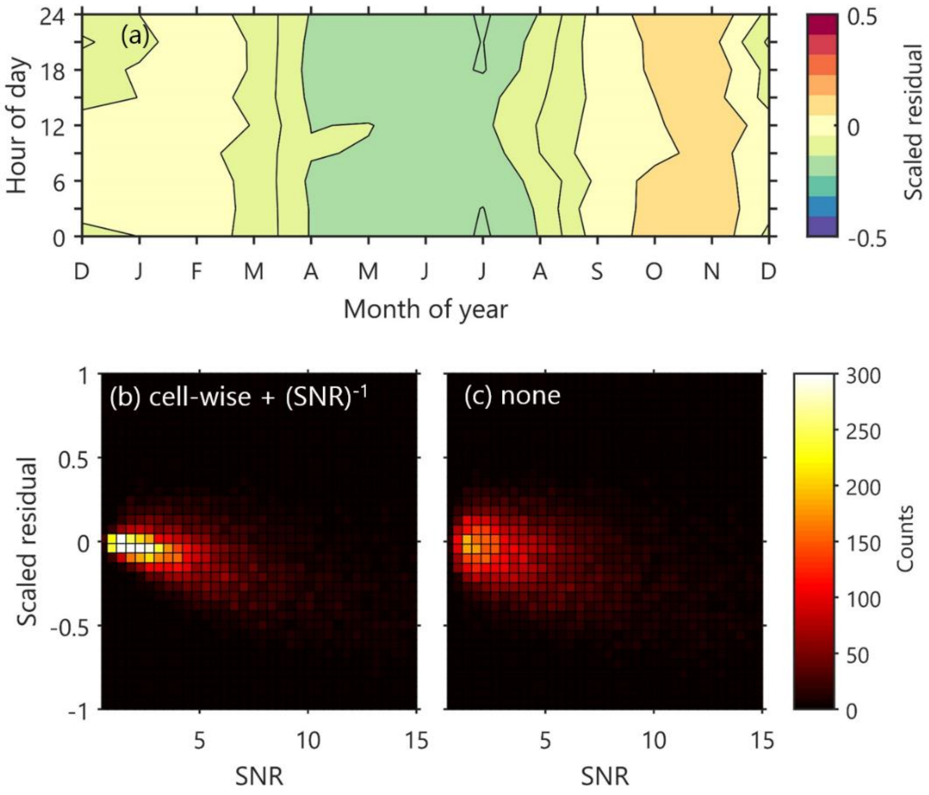

Figure S.4 The seasonal diurnal cycles of the scaled residual is visualised in panel (a) and 3d histograms of the scaled residual as a function of signal-to-noise (SNR) is presented in panels (b) and (c). The difference of panels (b) and (c) is that the weak variables were downweighted cell-wise using the (SNR)-1 method (Sec. 4.1.1) for panel (b) and panel (c) holds data from an rCMB run without downweighting bad or weak variables. As shown in panel (a) no clear diurnal patterns in the scaled residuals can be observed, but a seasonal cycle exists. As reflected in panels (b) and (c), this seasonality can be explained by the mean SNR, which is highest in summer when the OA mass loading is at highest (e.g. Heikkinen et al., 2020). The relationship between the scaled residual and SNR is highest when the errors have been further downweighted for rCMB to reduce the weight of weak and noisy variables within the PMF iterations.

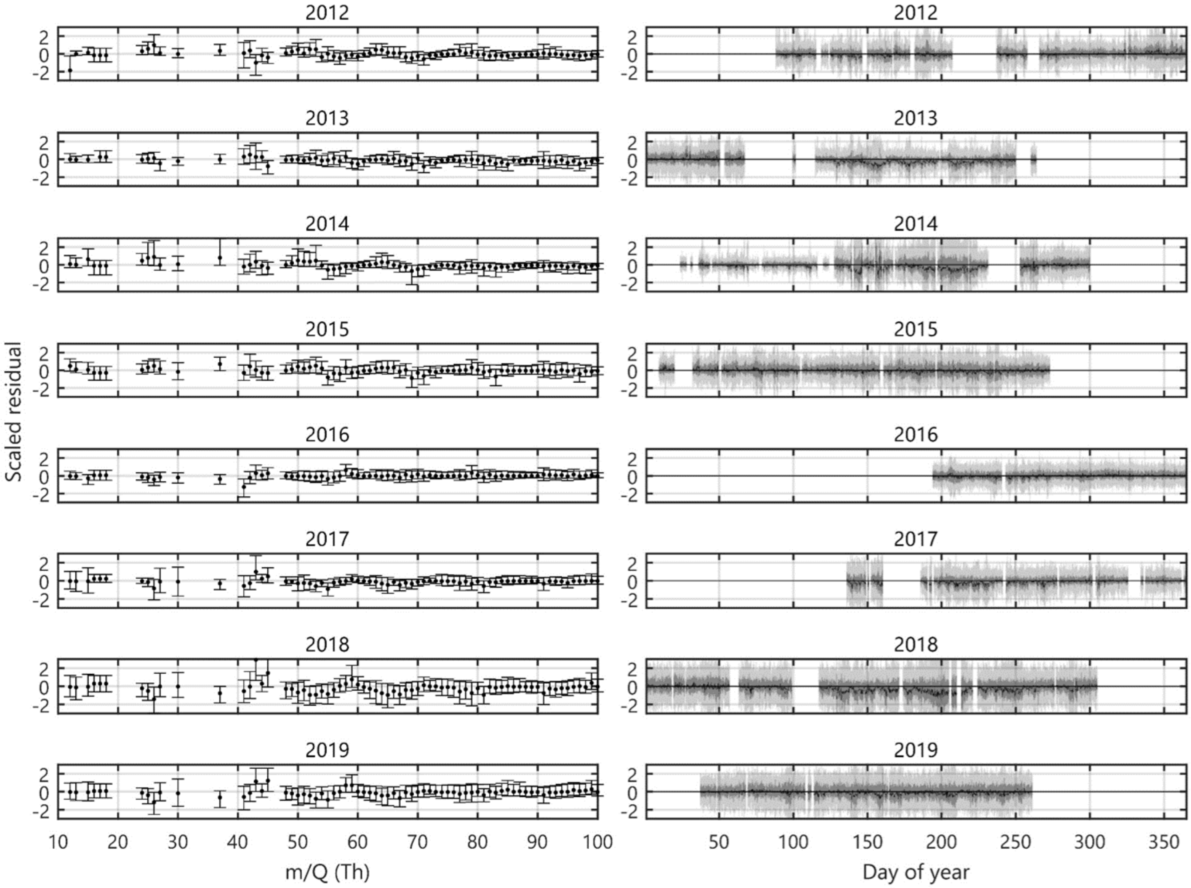

Figure S.5 The annual median (IQR shown with error bars) scaled residual spectra (left) and time series (90th and 10th percentiles with light grey shadings and IQR with darker grey shadings). Both show primarily noisy features. The slightly negative scaled residuals shown in figure 4 are visible during summers. These periods are associated with high SNR (i.e. summertime at SMEAR II).

40

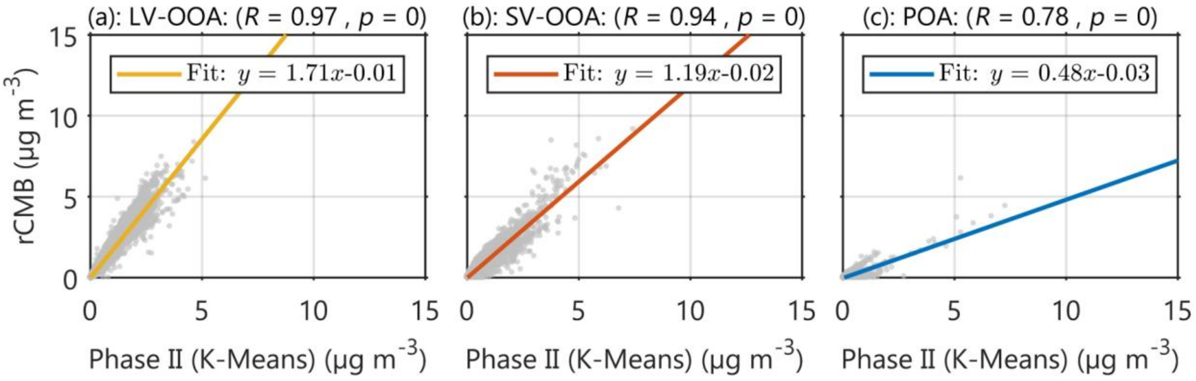

Figure S.6 Comparison of time series obtained after Phase II (K-Means) and rCMB for LV-OOA (panel a), SV-OOA (panel b) and POA (panel c). The coloured lines represent linear fits, and their equations are written in the panel captions. The panel titles contain Pearson correlation coefficient values (*R*) 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. The slope in panel (c) is highly driven by the high POA concentrations. These concentrations were higher within the initial free PMF compared to the rCMB run. We could suspect that the POA spectrum modelled by the initial PMF slightly differs from the rest of the POA spectra, which results a difference between the POA concentration between the rCMB and initial PMF.

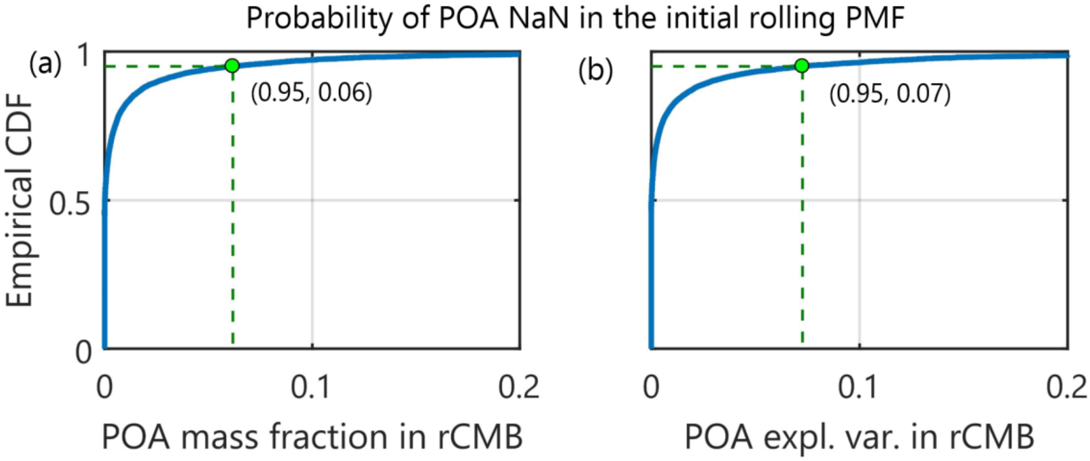

Figure S.7 The cumulative distribution functions showing the POA mass fraction calculated with rCMB when POA was not resolved in the unconstrained initial rolling PMF (i.e. when POA in the rolling window was NaN). The 95th percentile (3σ) is presented as the green marker. It corresponds to a POA mass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

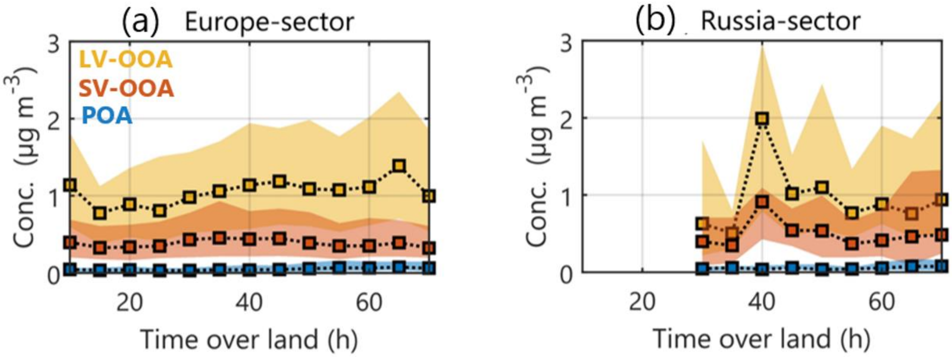

Figure S.8 The different rCMB factors (*y* axes in µg m^{-3}) vs TOL (*x* axes in hours) for the Europe-sector (panel a) and Russia-sector (panel b; see Fig. S.2 for a more precise sector definition). The data are binned to 5-hourly TOL bins. The shaded areas represents the concentration interquartile ranges (25th to 75th percentile) and the square markers the median concentrations.