Supplement of

Analysis of functional groups in atmospheric aerosols by infrared spectroscopy: method development for probabilistic modeling of organic carbon and organic matter concentrations

Charlotte Bürki[1], Matteo Reggente[1], Ann M. Dillner[2], Jenny L. Hand[3], Stephanie L. Shaw[4], and Satoshi Takahama[1]

[1]ENAC/IIE Swiss Federal Institute of Technology Lausanne (EFPL), Lausanne, CH-1015, Switzerland
[2]Air Quality Research Center, University of California Davis, Davis, CA 95616, USA
[3]Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, CO 80523, USA
[4]Electric Power Research Institute, Palo Alto, CA, 94304, United States

Correspondence: Satoshi Takahama (satoshi.takahama@epfl.ch)

Contents

S1 Contributions to the log-likelihood

In this section we outline calculations for assessing contribution of individual samples to the likelihood function (eq. 6). Let $r = (y - m_\mathrm{C})/\sigma$ represent the model residual normalized by the measurement precision. The contribution from a single sample to the overall likelihood $p(y|\theta) = \prod_{i \in \mathcal{S}} f_i$ is given by:

$$f_i = \left(\frac{1}{2\pi\sigma_i^2}\right)^{1/2} \exp\left[-\frac{1}{2}r_i^2\right]$$

Isolines of $\ln(f)$ (dropping the subscript i) can be generated (Figure S1) for several combinations of σ and r:

$$\ln f = -\frac{1}{2}\left[\ln(2\pi) + 2\ln(\sigma) + r^2\right] \ . \tag{S1}$$

This quantity gives an indication for the magnitude of contribution by individual data point (with uncertainty σ and relative deviation r) to the overall log-likelihood. For example, a sample near the detection limit ($m_\mathrm{C} \sim 3\sigma_0$) compared to one at the limit of quantification ($m_\mathrm{C} \sim 10\sigma_0$) means $\sigma = \sigma_0(1+3^2\kappa^2)^{1/2}$ and $\sigma_0(1+10^2\kappa^2)^{1/2}$, respectively, from our heteroscedastic error model (eq. 7). For identical r, the σ contribution of the higher concentration sample to $\ln f$ is ∼80% of the lower one for a for $\sigma_0 = 0.37$ µg cm^{-2} and $\kappa = 0.07$ (Section 3.2) but decreases to ∼25% for $\kappa = 0.3$.

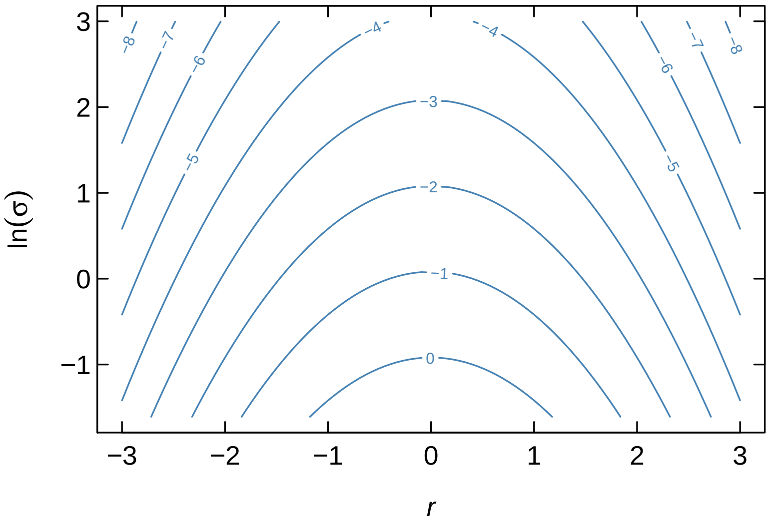

Figure S1. Isolines of $\ln f$ according to eq. S1.

S2 Prior distributions

This section includes Figures S2–S5; and Table S1.

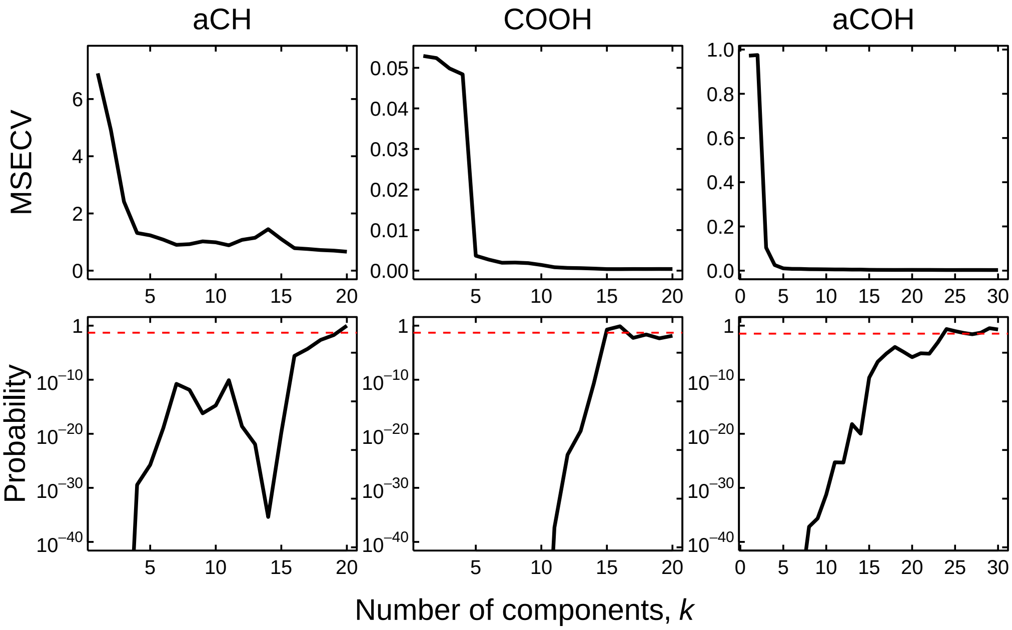

Figure S2. MSECV curves (in units of μmole of FG, top row) and resulting prior probability distributions for k (bottom row). Horizontal lines in in bottom row correspond to probability for a uniform distribution over the selected number of components.

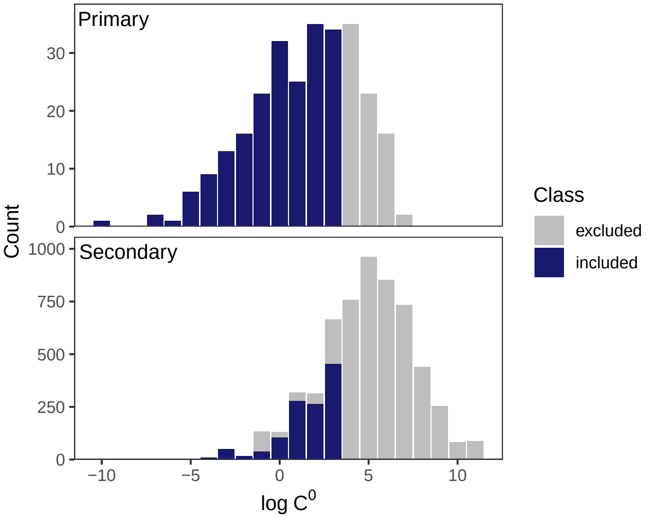

Figure S3. Distribution of equilibrium vapor concentrations C^0 ($\mu\mathrm{g\,m^{-3}}$) for molecules taken from Rogge et al. (1993) and Rogge et al. (1998) ("Primary") and the MCM v3.3.1 database (Jenkin et al., 1997; Saunders et al., 2003) ("Secondary"). Only non-radical molecules with $C^0 \leq 10^{3.5}$ $\mu\mathrm{g\,m^{-3}}$ are used in this study (excluded molecules below this threshold in the "Secondary" category represent radical species).

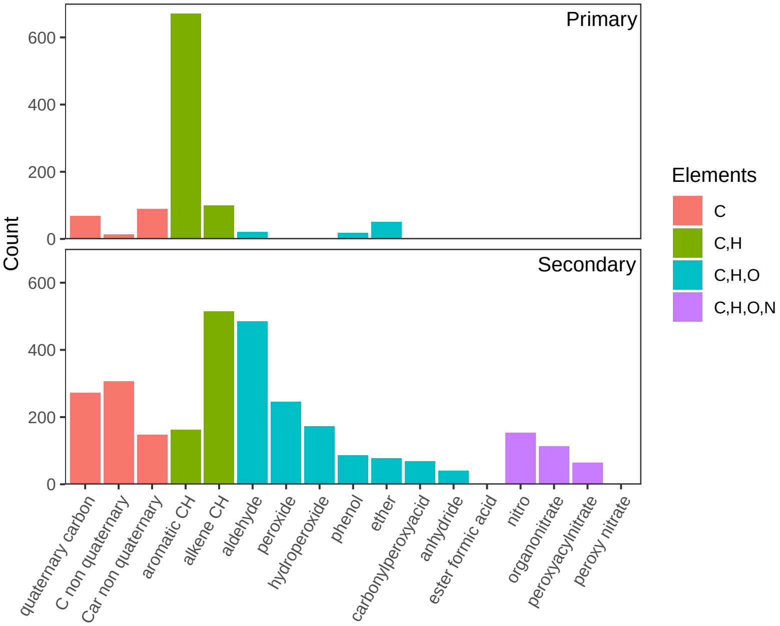

Figure S4. Number of molecular structures associated with undetected carbon atoms for all semivolatile compounds selected in Figure S3. Structures are colored by the elements that they contain. Structure names are described with illustrations in Table 1 of technical note by Ruggeri and Takahama (2016).

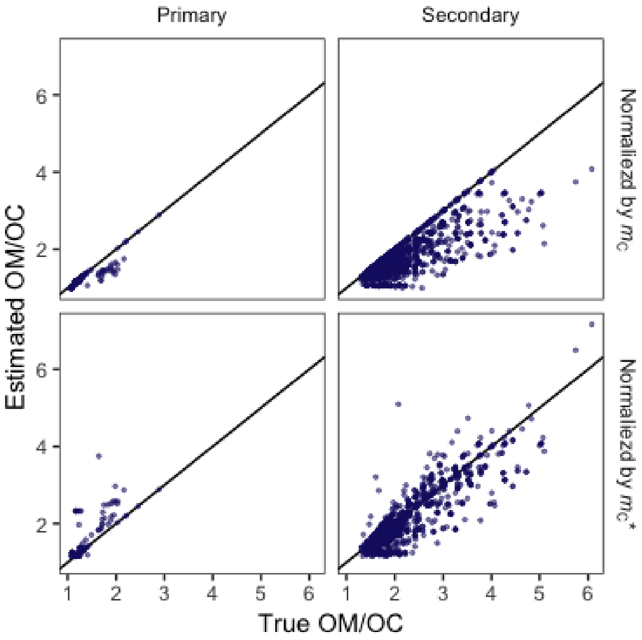

Figure S5. Estimates of OM/OC when normalized by m_{C} and αm_{C}. Secondary aerosol species contain many small but highly functional molecules, but the overall mode of the true OM/OC distribution is 1.96; the mode for primary aerosol species is 1.17.

Table S1. Average number of atoms attached to each type of bond assumed for various types of mixtures. $\lambda_{\mathrm{C,COOH}} = \lambda_{\mathrm{C,carbonyl}} = 1$. Table adapted from Takahama and Ruggeri (2017).

Study	Mixture type	$\lambda_{\mathrm{C,CH}}$	$\lambda_{\mathrm{C,aCOH}}$
Allen et al. (1994)	ambient	0.5	
Russell (2003)	ambient	0.5	1
Reff et al. (2007)	indoor/ambient	0.48	
Chhabra et al. (2011)	α-pinene SOA	0.63	0.63
	guaiacol SOA	0.88	0.88
Russell and co-workers*	ambient	0.5	0.5
Ruthenburg et al. (2014)	ambient	0.5	0
Takahama and Ruggeri (2017)**	α-pinene SOA	0.39–0.5	0.09–0.52

*reflects assumptions by Russell et al. (2009), Liu et al. (2009), and Day et al. (2010).

**estimated from simulated molecular mixtures.

S3 Cluster analysis

This section includes Figures S6–S9.

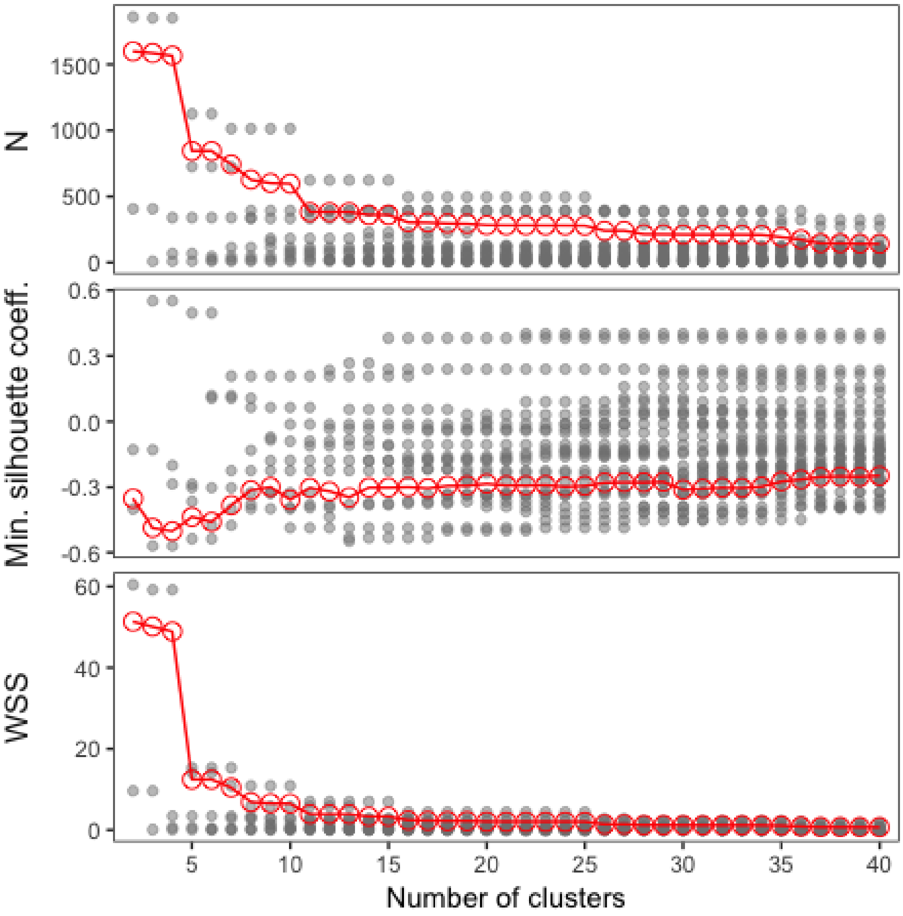

Figure S6. Number of samples in each cluster, minimum silhouette coefficient, and within sum-of-squares of each cluster as a function of the number of clusters formed. Gray points represent individual clusters, and red points and lines are values averaged across clusters.

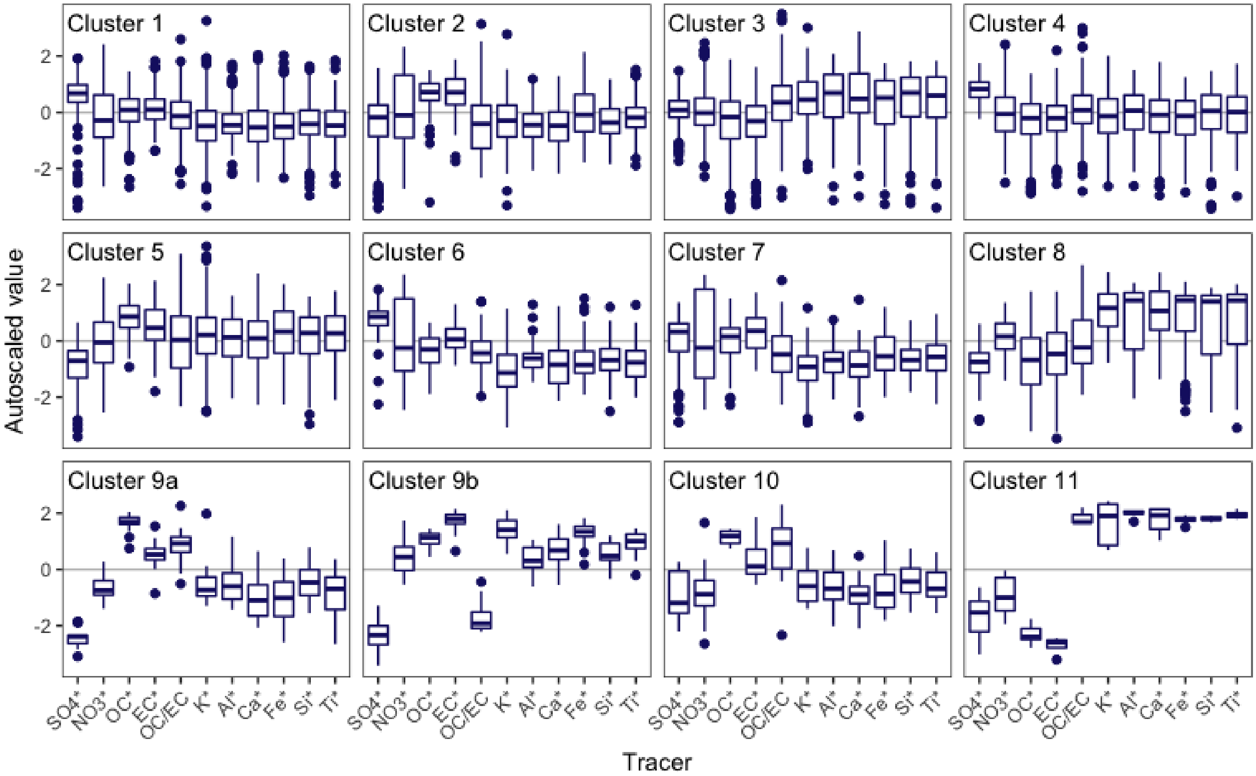

Figure S7. Comparisons of relative tracer concentrations. "*" denotes $PM_{2.5}$-normalized quantities. Normalized values are first logarithmically-transformed to be approximately symmetric, and then autoscaled (mean-centered and normalized by standard deviation of the variable for the entire data set). Values greater than zero for a particular cluster indicates that this substance or ratio is enriched in samples belonging to this cluster, relative to the rest of the samples.

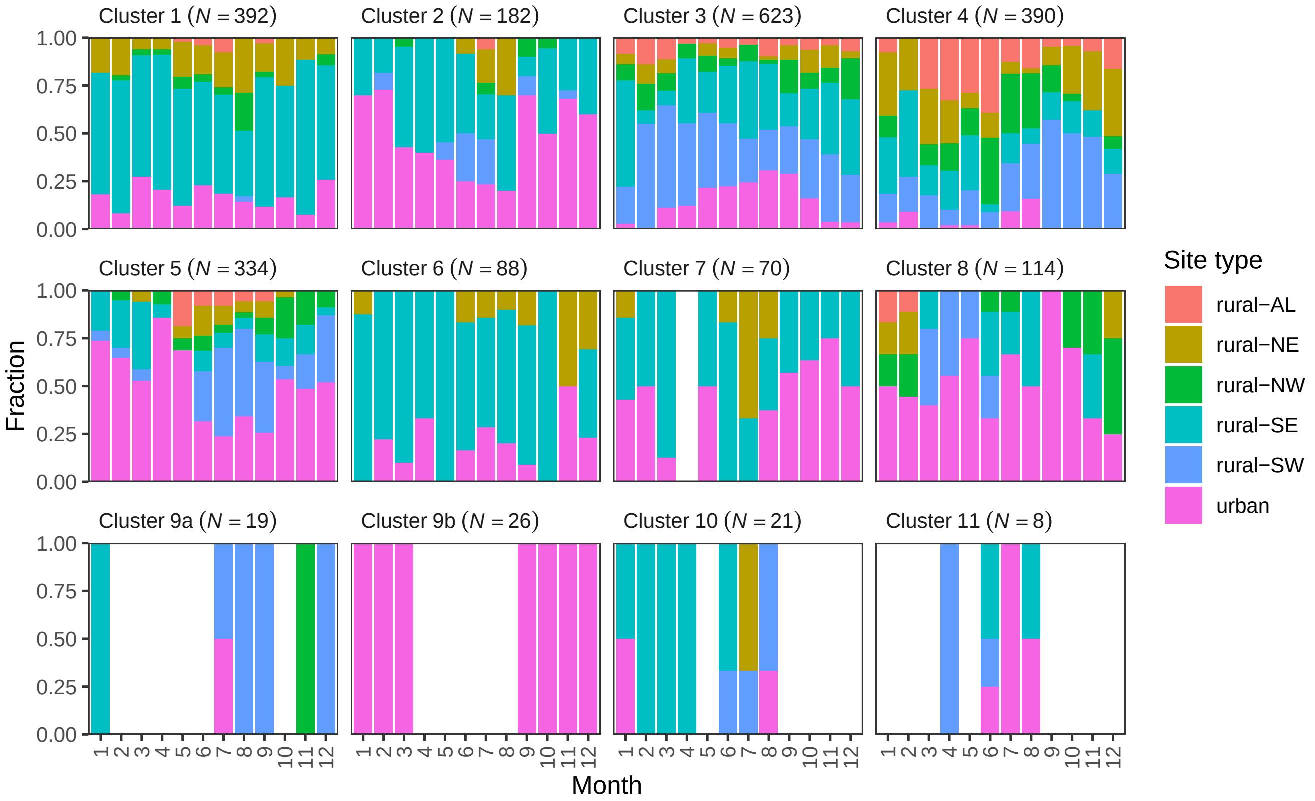

Figure S8. Composition of clusters.

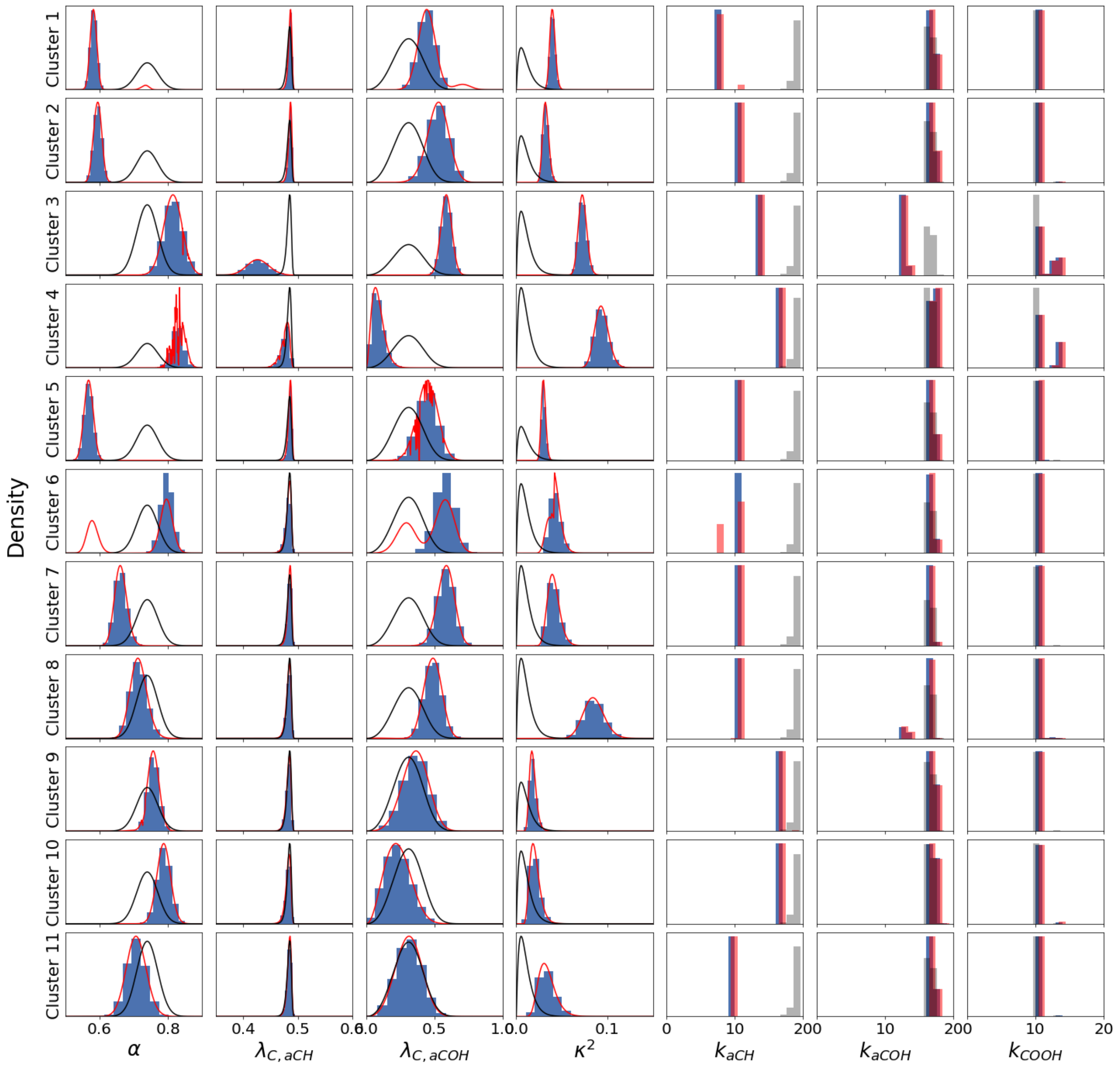

Figure S9. Posterior distributions of parameters for each cluster from MCMC (blue histograms) and L-BFGS-B (red lines) compared to prior distributions (black lines).

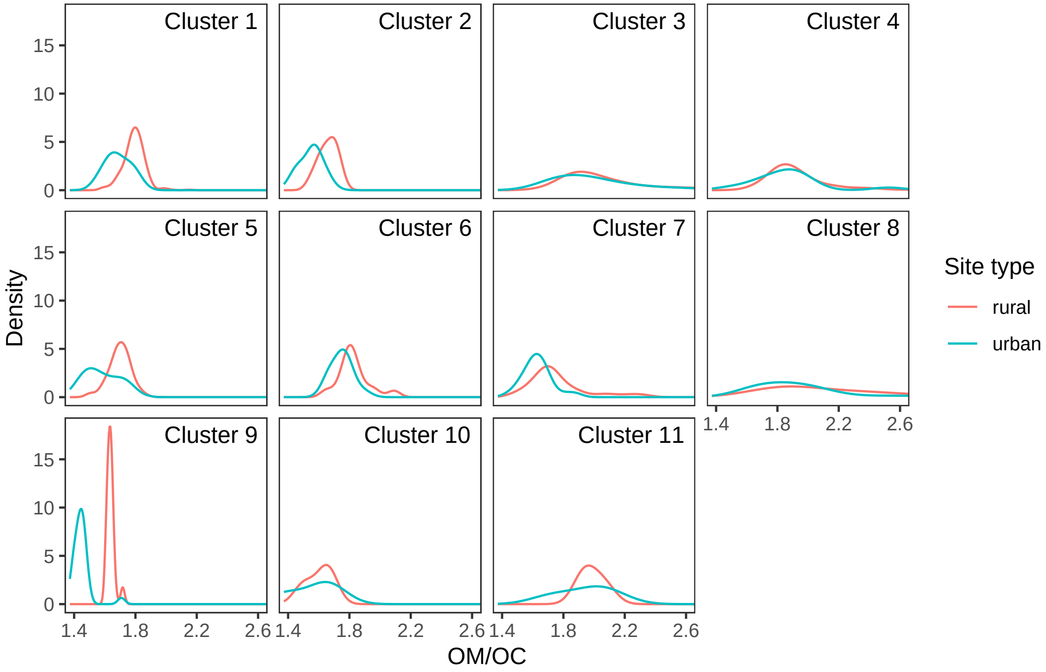

Figure S10. Probability distributions of OM/OC ratios segregated by site type.

S4 Posterior predictions

After obtaining the posterior parameter distributions, probability distributions and intervals of predictions of the target variable y are obtained for model checking (Robert, 2007; Vehtari and Ojanen, 2012; Gelman et al., 2013). The posterior predictive distribution for new $\tilde{y}$ from spectrum $\tilde{x}$ is given by

$$p(\tilde{y}|y) = \int_{\theta} p(\tilde{y}|\theta) p(\theta|y) d\theta \,. \tag{S2}$$

For model checking, $\tilde{y}$ corresponds to replications of the data used for fitting; the integral in eq. S2 can be numerically evaluated using the values of θ generated from MCMC. The expected value of this posterior distribution corresponds to m_{C} (eq. 5). While m_{C} is uniquely determined for a given realization of θ, ε and therefore $\tilde{y}$ varies according to the sample drawn from a normal distribution characterized with the value of κ^2. The posterior predictive distribution is generally symmetric, and the mode or mean of $\tilde{y}$ can simply be approximated by the mode or mean of the posterior parameter distributions (Figure S11).

More generally, for any scalar-valued property z (e.g., m_{C} or OM/OC) dependent on $\psi = \theta \setminus \{\kappa^2\}$, $p(z|y)$ and its corresponding central estimate or intervals can also be constructed by transforming the Markov sequence of the parameters: $\{z(\psi^{[1]}), z(\psi^{[2]}), \ldots, z(\psi^{[n]})\}$ (Hoff, 2009). In applying this strategy toward the calculation of OM/OC ratios, we obtain posterior probability distributions for each sample. Due to the nonzero probabilities of several discrete values of k_{aCOH} and k_{COOH}, OM/OC estimates can become multimodal when contributions from these oxygenated FGs are substantial (examples shown in Figure S12). We find that the median or peak of the largest mode of the posterior distribution of OM/OC is well-approximated by the maximum a posteriori estimate (MAP; Section D) of the parameters (slope and correlation coefficient of 1.0) and so we report this value as the single-point estimate of OM/OC for each sample. The span of 95% prediction intervals (representing uncertainties in sample-specific OM/OC values due to uncertainties in FG-OC model parameters) generally corresponds to less than 6% the reported OM/OC for most samples, except for clusters 8 and 11 where many samples had interval spans extending up to 20 and 10% of the value of the mode, respectively. Cluster 8 had larger intervals due to the two noncontiguous sets of k_{aCOH} with substantial probabilities, leading to separation in the modes of OM/OC. In such instances these samples, may benefit from further disaggregation for parameter estimation or incorporation of observations more specific toward the oxygenated fraction to reduce posterior parameter uncertainties. The high uncertainty in prediction for samples in cluster 11 is due to the small number ($N = 8$) samples in this cluster, resulting in broad posterior parameter distributions. Hierarchical Bayesian modeling (Gelman and Hill, 2007) may be beneficial in leveraging relationship of small subgroups of samples to the greater population to better handle such cases.

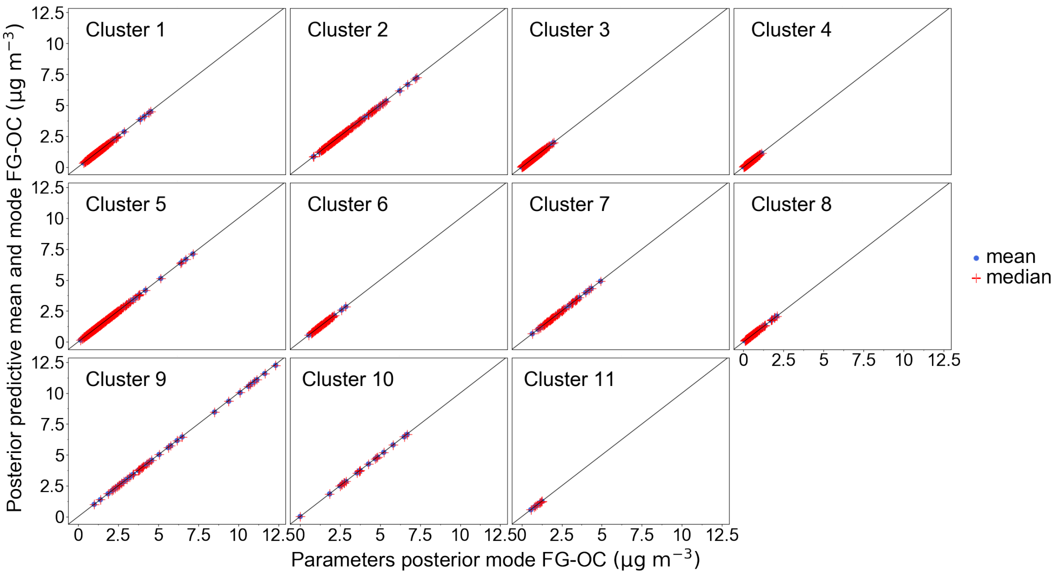

Figure S11. Comparison of central values of the posterior predictive distribution with predictions from single-point estimates of parameters obtained from their respective distributions.

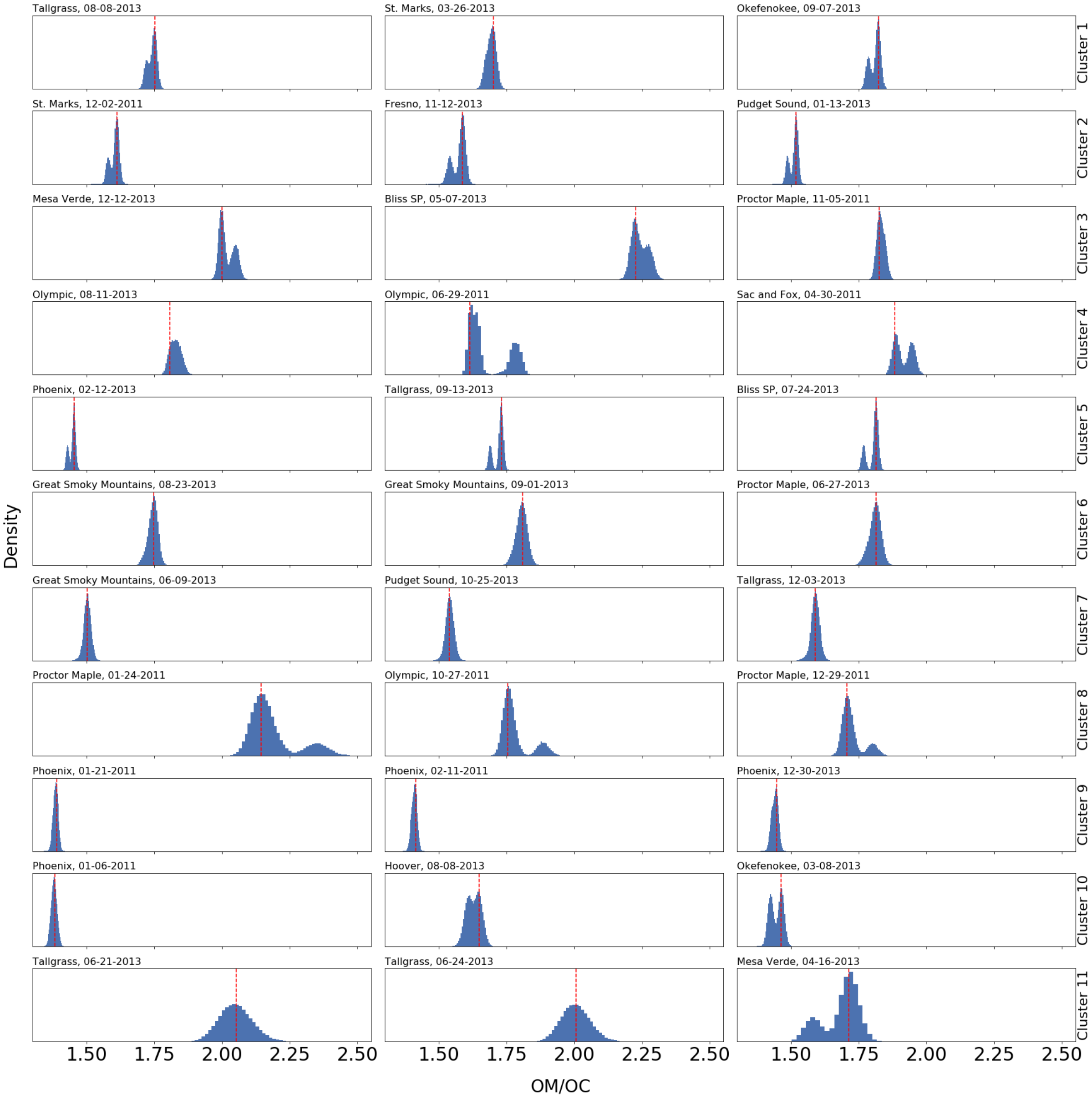

Figure S12. Posterior distributions for OM/OC. Red vertical lines indicate the reported value using MAP estimates of the parameters.

S5 Spatial and temporal prevalence of cluster types

45 This section includes Figures S13 and S14.

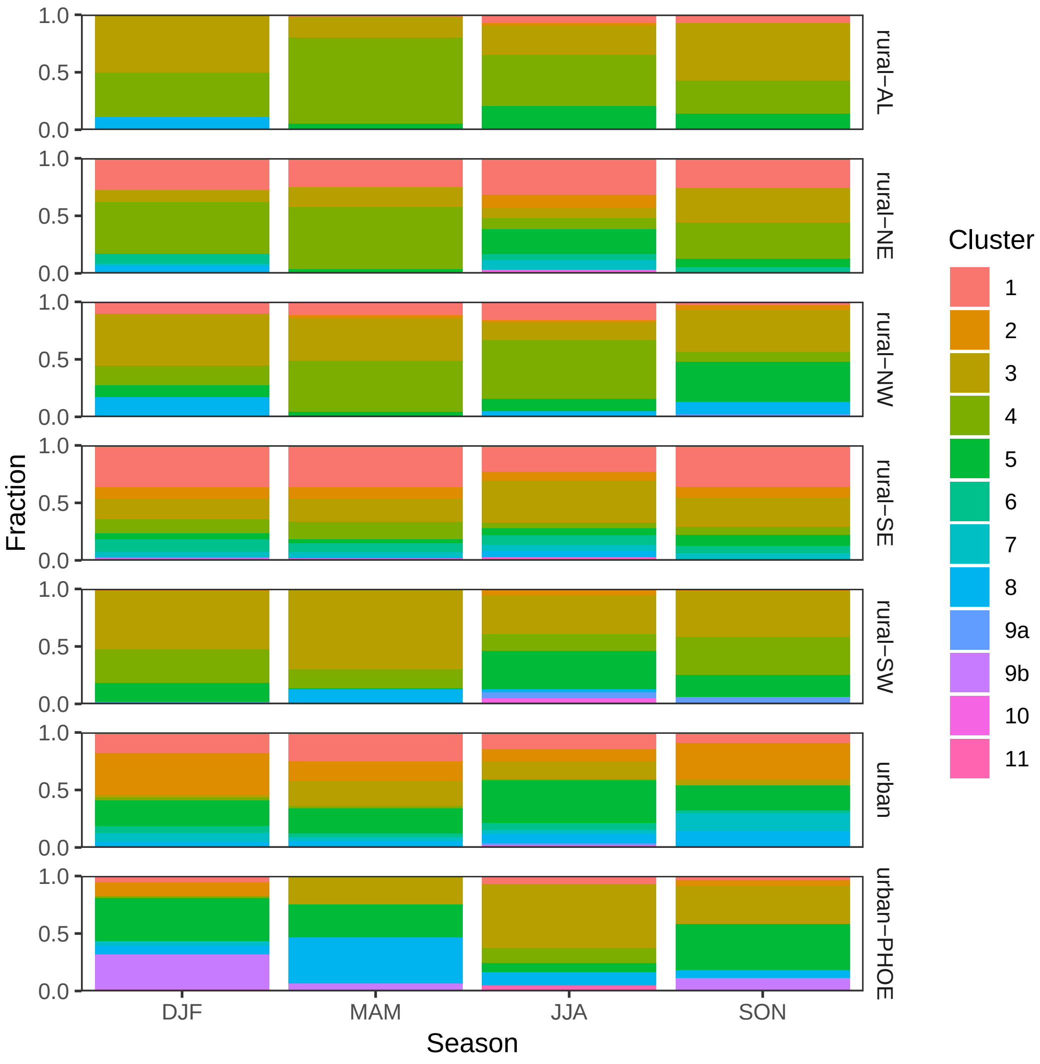

Figure S13. Frequency of clusters as fraction of samples at each site and season. “urban-PHOE” refers to Phoenix, AZ, and “urban” refers to all other urban sites.

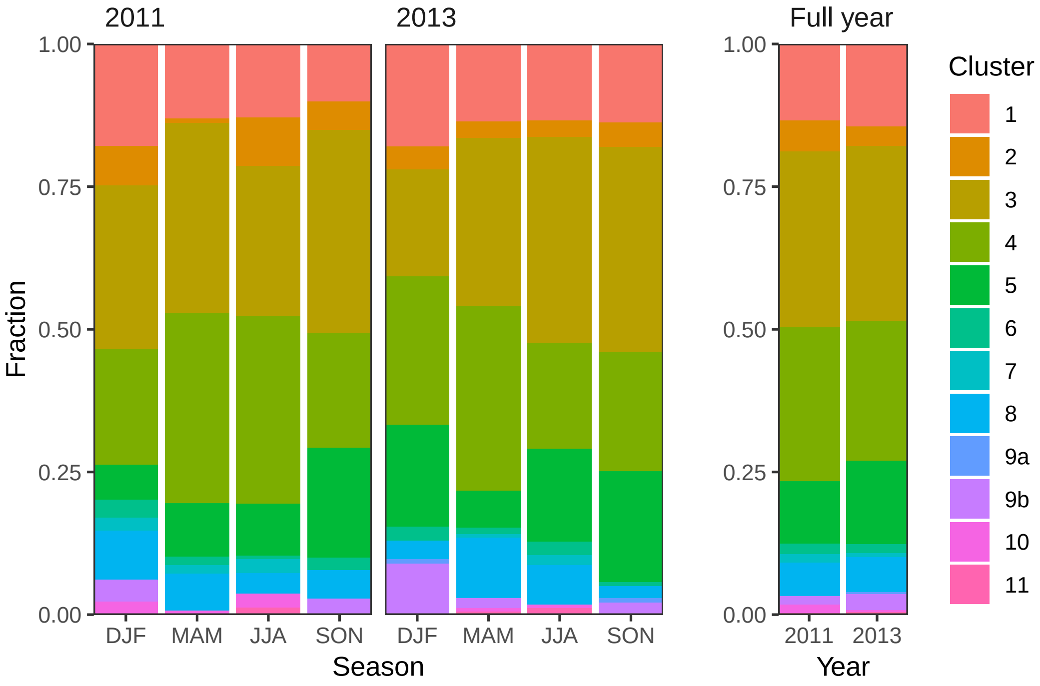

Figure S14. Frequency of clusters as fraction of samples during each year and season for six sites.

References

Allen, D. T., Palen, E. J., Haimov, M. I., Hering, S. V., and Young, J. R.: Fourier-transform Infrared-spectroscopy of Aerosol Collected In A Low-pressure Impactor (LPI/FTIR) - Method Development and Field Calibration, Aerosol Science and Technology, 21, 325–342, https://doi.org/10.1080/02786829408959719, 1994.

Chhabra, P. S., Ng, N. L., Canagaratna, M. R., Corrigan, A. L., Russell, L. M., Worsnop, D. R., Flagan, R. C., and Seinfeld, J. H.: Elemental composition and oxidation of chamber organic aerosol, Atmospheric Chemistry and Physics, 11, 8827–8845, https://doi.org/10.5194/acp-11-8827-2011, http://www.atmos-chem-phys.net/11/8827/2011/, 2011.

Day, D. A., Liu, S., Russell, L. M., and Ziemann, P. J.: Organonitrate group concentrations in submicron particles with high nitrate and organic fractions in coastal southern California, Atmospheric Environment, 44, 1970–1979, https://doi.org/10.1016/j.atmosenv.2010.02.045, 2010.

Gelman, A. and Hill, J.: Data Analysis Using Regression and Multileve//Hierarchical Models, Cambridge Univ. Press, Cambridge, 2007.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D.: Bayesian Data Analysis, Chapman & Hall/CRC Texts in Statistical Science, Chapman & Hall/CRC, New York, NY, 3rd edn., 2013.

Hoff, P. D.: A First Course in Bayesian Statistical Methods, Springer, New York, NY, https://doi.org/10.1007/978-0-387-92407-6, 2009.

Jenkin, M. E., Saunders, S. M., and Pilling, M. J.: The tropospheric degradation of volatile organic compounds: a protocol for mechanism development, Atmospheric Environment, 31, 81–104, https://doi.org/10.1016/S1352-2310(96)00105-7, http://www.sciencedirect.com/science/article/pii/S1352231096001057, 1997.

Liu, S., Takahama, S., Russell, L. M., Gilardoni, S., and Baumgardner, D.: Oxygenated organic functional groups and their sources in single and submicron organic particles in MILAGRO 2006 campaign, Atmospheric Chemistry and Physics, 9, 6849–6863, https://doi.org/10.5194/acp-9-6849-2009, 2009.

Reff, A., Turpin, B. J., Offenberg, J. H., Weisel, C. P., Zhang, J., Morandi, M., Stock, T., Colome, S., and Winer, A.: A functional group characterization of organic PM2.5 exposure: Results from the RIOPA study RID C-3787-2009, Atmospheric Environment, 41, 4585–4598, https://doi.org/10.1016/j.atmosenv.2007.03.054, 2007.

Robert, C. P.: The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation, Springer Texts in Statistics, Springer, New York, NY, 2nd edn., 2007.

Rogge, W. F., Hildemann, L. M., Mazurek, M. A., Cass, G. R., and Simoneit, B. R. T.: Sources of Fine Organic Aerosol .2. Noncatalyst and Catalyst-equipped Automobiles and Heavy-duty Diesel Trucks, Environmental Science & Technology, 27, 636–651, https://doi.org/10.1021/es00041a007, 1993.

Rogge, W. F., Hildemann, L. M., Mazurek, M. A., Cass, G. R., and Simoneit, B. R. T.: Sources of fine organic aerosol. 9. Pine, oak and synthetic log combustion in residential fireplaces, Environmental Science & Technology, 32, 13–22, https://doi.org/10.1021/es960930b, 1998.

Ruggeri, G. and Takahama, S.: Technical Note: Development of chemoinformatic tools to enumerate functional groups in molecules for organic aerosol characterization, Atmospheric Chemistry and Physics, 16, 4401–4422, https://doi.org/10.5194/acp-16-4401-2016, http://www.atmos-chem-phys.net/16/4401/2016/, 2016.

Russell, L. M.: Aerosol organic-mass-to-organic-carbon ratio measurements, Environmental Science & Technology, 37, 2982–2987, https://doi.org/10.1021/es026123w, 2003.

Russell, L. M., Bahadur, R., Hawkins, L. N., Allan, J., Baumgardner, D., Quinn, P. K., and Bates, T. S.: Organic aerosol characterization by complementary measurements of chemical bonds and molecular fragments, Atmospheric Environment, 43, 6100–6105, https://doi.org/10.1016/j.atmosenv.2009.09.036, 2009.

Ruthenburg, T. C., Perlin, P. C., Liu, V., McDade, C. E., and Dillner, A. M.: Determination of organic matter and organic matter to organic carbon ratios by infrared spectroscopy with application to selected sites in the IMPROVE network, Atmospheric Environment, 86, 47–57, https://doi.org/10.1016/j.atmosenv.2013.12.034, 2014.

Saunders, S. M., Jenkin, M. E., Derwent, R. G., and Pilling, M. J.: Protocol for the development of the Master Chemical Mechanism, MCM v3 (Part A): tropospheric degradation of non-aromatic volatile organic compounds, Atmospheric Chemistry and Physics, 3, 161–180, https://doi.org/10.5194/acp-3-161-2003, 2003.

Takahama, S. and Ruggeri, G.: Technical note: Relating functional group measurements to carbon types for improved model–measurement comparisons of organic aerosol composition, Atmospheric Chemistry and Physics, 17, 4433–4450, https://doi.org/10.5194/acp-17-4433-2017, https://www.atmos-chem-phys.net/17/4433/2017/, 2017.

Vehtari, A. and Ojanen, J.: A survey of Bayesian predictive methods for model assessment, selection and comparison, Statist. Surv., 6, 142–228, https://doi.org/10.1214/12-SS102, https://doi.org/10.1214/12-SS102, 2012.