Response to reviewer comments for manuscript: "Analysis of functional groups in atmospheric aerosols by infrared spectroscopy: method development for probabilistic modeling of organic carbon and organic matter concentrations"

Reviewer 1

The manuscript describes a statistical model and a probabilistic framework to characterize combinations of organic matter, organic carbon and functional groups obtained from the Fourier transform infrared (FTIR) spectra of fine particulate matter (PM2.5). The model was found to be consistent with field measurements of organic carbon (OC). The Development of these models and frameworks is important and timely as they can be used in developing machine learning algorithms.

We thank the reviewer for the encouraging assessment.

1. The abstract is too long. Consider shortening.

   We have substantially shortened the abstract to highlight the essential points of the manuscript.

2. The introduction is missing a discussion on the results obtained from NMR spectroscopy in the characterizing organic aerosols and using multivariate analysis to correlate functional groups with sources of aerosols. For example, see Chapter Two - NMR Studies of Organic Aerosols, Annual Reports on NMR Spectroscopy, Volume 92, 2017, Pages 83-135.

   We thank the reviewer for pointing out this useful review. In current form it is difficult to relate our results to that reported by NMR studies given several unknowns: i) systematic differences in composition between the water soluble fraction (targeted by NMR) and $PM_{2.5}$ analyzed by FTIR in this work, ii) sampling artifacts related to high and low mass loadings (e.g., gas/particle partitioning), and iii) the different scales of functional group ratios used by the two techniques. We hope to dedicate a future study (or series of studies) relating the two techniques for source identification or source apportionment by receptor modeling.

   We have included a citation to the work to indicate its relevance for future comparisons.

3. Figure 4: this figure has 12 sub figures; each is labeled with a 'cluster' number and contain a number of overlapping spectra as judged by the gradation in the line colors. The caption needs to be modified to refer the reader to appropriate table or graph containing the definition of each cluster.

   We have now specified in the caption: "The clustering procedure and interpretation are described in Sections 3.1 and 4.2, respectively."

4. Figure 7: The caption needs to be modified to refer the reader to appropriate table of graph containing the definition of parameters in the y- and x-axes.

   We have now specified in the caption:" "Density" refers to the probability or mass density and the variables are described in Sections 1.1 and 3.2."

5. Figure 8, 9, 10, 11: the use of three-letter abbreviation for seasons that do not correspond to the actual name of the season is confusing: DJF = winter, MAM = spring, JJA = summer, and SON = fall. It is better to use WR = winter, SG = spring, SR = summer, and FL = fall.

   We thank the reviewer for the suggestion, but DFJ, MAM, JJA, SON is a standard convention for denoting seasons — we have now written out the name of the months that they stand for in the caption of Figure 8: "X-axes denote seasons: DJF (December, January, February) = winter, MAM (March, April, May) = spring, JJA (June, July, August) = summer, and SON (September, October, November) = fall."

Reviewer 2

The manuscript describes a novel approach using FT-IR and functional groups to obtain organic matter to organic carbon ratios. The manuscript fits the scope of Atmospheric Measurement Techniques. The data and approach are novel and appear of good quality. The manuscript could benefit from some attention to detail and some mostly minor clarifications.

We thank the reviewer for the encouraging assessment.

1. *Regarding the FTIR technique*:

   - A fundamental question that I do not see mentioned or addressed is that ATR spectroscopy is highly sensitive to the deposit structure and depth homogeneity. I am missing a clear discussion on how the FG approach compares with PM mass? i.e. if the data gets less consistent with TOR when there is more deposit on the filter? Or not? This seems a critical thing to discuss?
   - A second related issue is what happens when diurnally different sources are important? E.g. vehicles AM and wood burning PM as the corresponding molecules are now in different layers and as stated above ATR is sensitive to penetration depth with higher sensitivity to molecules closer to the surface. Please comment? Could Phoenix data show something there?

   We regret having omitted the details of the analysis that has caused understandable confusion.

   One of the coauthors previously used ATR-FTIR to analyze PTFE filters (Coury and Dillner, 2008), which is indeed sensitive to penetration depth and particle morphology. Some of these influences on apparent absorbances of particle samples are described by other publications, including ours (Kortüm, 1969; Harrick, 1979; Milosevic, 2012; Arangio et al., 2019).

   For this work, we have used transmission mode analysis as described by Maria et al. (2003) and Ruthenburg et al. (2014). The samples collected in this way are optically thin and so penetration depth is not an issue (transmittance does not reach 0%) (Maria et al., 2003), which is a common concern with transmission mode analysis. Diurnal layering of different particle types can lead to a nonuniform film structure, but empirically we find that the measured transmittance is within measurement error (Debus et al., 2019) regardless of which face of the filter is directed toward the incident beam. We therefore assume that this does not have a detectable impact on our interpretations.

   While a "film" of submicron particles analyzed by transmission mode can be modeled as a substance with a single, effective refractive index (Fischer, 1975; Choy, 2016), presence of larger particles can potentially lead to i) scattering and ii) anomalous dispersion (Leisner and Wagner, 2010). The former manifests itself in reduced transmittance and therefore increased apparent absorbance (e.g., Dazzi et al., 2013). This is handled by baseline correction, which removes the scattering contribution from both the PTFE substrate and also the particles

(Takahama et al., 2019). Samples with anomalous dispersion are found in the "anomalous" clusters in this work (referring the reader to Reggente et al., 2019, in which further details are provided), and caveats are provided for their interpretation. Furthermore, the additive mixing model of Bougher-Lambert-Beer (BLB) assumes dilute, non-interacting mixtures, which may not strictly apply in our samples. However, the "inverse calibration" approach of eq. 2 is robust with respect to a certain degree of nonlinearity in the data (Griffiths and Haseth, 2007). These issues would effect the calibration which estimates the molar density n (eq. 2), but not for the overall framework of the FG-OC model (eqs. 1 and 5). The FG-OC model is, in principle, agnostic with respect to the method by which n is estimated.

A separate manuscript describing particle models for mid-infrared analysis has been in preparation and the topics mentioned above will be explored in more depth. For this manuscript, we have added the following statement in Section 2:

"PTFE of ambient and laboratory samples were analyzed nondestructively by FTIR in transmission mode (Maria et al., 2003) after placing them in a custom minichamber purged with air passed through a molecular sieve to remove water vapor and carbon dioxide (Ruthenburg et al., 2014; Debus et al., 2019). Spectra were truncated to the region above 1500 cm^{-1} and baseline corrected (Kuzmiakova et al., 2016) to reduce scattering contributions from the PTFE filter (McClenny et al., 1985) and particles (Takahama et al., 2019). Further details on the sample collection, analysis, and spectra processing steps are described by previous works (Ruthenburg et al., 2014; Reggente et al., 2016; Debus et al., 2019; Takahama et al., 2019)."

2. Phoenix is being highlighted in the manuscript, which surprised me as I thought of IMPROVE being a rural network. In any case, even more surprising to me is the Phoenix woodburning. Can you please provide peer reviewed literature that actually supports that biomass burning is so dominant over the whole wintertime (3 months, if they have even 3 months of winter in Phoenix) and all this is not just some mass/deposit heterogeneity artifact as questioned above related to very strong inversion periods?

   We thank the reviewer for pointing out the absence of proper citations which we have now added to Section 4.2. IMPROVE does contain a few urban sites; wood burning in Phoenix, AZ, has previously reported by Ramadan et al. (2011) and the "official fireplace season" lasts through October through February (Pope et al., 2017). In Cluster 9b, 22 out of the 26 samples fall in this period, with the remaining samples coming from September (1 sample) and March (3 samples).

3. The manuscript could benefit from more attention to detail. Examples: L14 and throughout the manuscript-Please use subscripts in your molecule numbers. L14 COOH would be carboxylic acid especially when contrasted with carboxylate.

   "carboxylic COO" should have been written "carboxylic COOH" on L14 and we apologize for this error. But in general, there is no molecule number because we use the approach of calibrating to functional groups rather than individual molecules (Anderson and Seyfried, 1948; Allen et al., 1994; Russell, 2003). We have now made it clear in the abstract that we are using functional group calibrations.

4. Figure 3: One molecule (dione) has carbons with 5 bonds (get rid of that double bond in the keto containing cycle). Also for all these species provide the correct names; not only the random MCM naming. The second molecule is poorly cropped and cut.

   We thank the reviewer for pointing out these errors. We have revised the figure and included standard names for all molecules shown.

5. Other comments:

- L19 R2 values are meaningless without n? or discussion of statistical relevance.

  We did not intend to imply that the agreement was perfect. We have now stated "the resulting calibrations reproduce TOR OC concentrations with reasonable agreement (r = 0.96 for 2474 samples) and provide OM/OC values generally consistent with our current best estimate of ambient OC." We had also mistakenly written the value of Pearson's correlation coefficient (r) as the coefficient of determination (R^2) for OLS regression so this has also been corrected.

- What is the "Reconstructed Fine Mass Equation", I assume this is some American network thing?

  This is an equation of mass closure defined by the IMPROVE network that expresses mass closure between gravimetric $PM_{2.5}$ mass and its major mass constituents. We have included this statement in Section 3.3:

  "For comparison, we estimate OM/OC as interpreted by coefficients of the RCFM equation (a statement of mass closure) [...]"

- (L34) Blanks: why was only ammonium sulfate used as blanks? Not ammonium nitrate or other materials?

  For a limited set of resources there is always a tradeoff between sampling more organic compounds versus non-organic blanks. With regards to ammonium nitrate specifically, including it in our calibrations would not necessarily introduce additional information as we have only used the region of FTIR spectra above 1500 cm^{-1}, where the N-H stretch in the ammonium is already captured by the ammonium sulfate.

  The effect of including ammonium nitrate, water, and other non-organic interferents in PLS calibrations has been explored by Boris et al. (2019), who found that organic functional group abundances could be predicted with similar accuracy but with possible differences in the overall model complexity (i.e., number of latent variables). While such conclusions likely depend on the types and proportion of calibration and blank samples, in our calibration set (as well as those of Boris et al., 2019), organic compounds not containing the functional group being calibrated also serve as additional blanks. Therefore, our calibration models are not deficient in blanks for each functional group, though this is an area that can be further explored. As described in the manuscript ("the framework is described generally such that it can [...] systematically evaluate improvements in calibrations with new standards or FGs"), the Bayesian framework presented here provides a means to incorporate new information and evaluate its value based on how it affects parameter posterior distributions.

- Figure 2, could you provide quality parameters on the fits?

  The chi-square statistic for all fits are greater than 300 (normally 1 is considered a "good fit"). Nonparametric models could be used to achieve more precise representations of the generated distributions, but it is not meant to be overly precise since the distribution-generating processes itself is based on approximations — they are dependent on the molecules available in the database, subsets used for estimation (Appendix C), and so on. Parametric models are therefore used to characterize the main features of these distributions (e.g., centrality, dispersion, and symmetry).

- Figure 5: Could you show a trendline with equation here? To get a quantitative idea of the drift?

  We have added a trendline with fit parameters in the caption.

- Figure 9: Probability density function are often given in units that if the curve gets integrated

it equals 1. In this manuscript, it seems never normalized and the density always arbitrary? In Fig 9 on some panels one wonders if all areas under the curve are identical though?

The probability densities are approximated by kernel density estimation (Hastie et al., 2009) (eq. 1), which should integrate to one. Defining a transformed variable $u = (x - x_j)/b$ such that $dx = -b\,du$, each local kernel K with bandwidth b evaluated along x_j for $j = \{1, 2, \ldots, N\}$ integrates to unity (eq. 2); and the resulting integral of the kernel density estimate of p can also be shown to be unity (eq. 3).

$$p(x) = \frac{1}{Nb} \sum_{j=1}^{N} K\left(\frac{x - x_j}{b}\right) \tag{1}$$

$$\int_{-\infty}^{\infty} K(u)du = 1 \tag{2}$$

$$\int_{-\infty}^{\infty} p(x)dx = 1 \tag{3}$$

Some curves may not strictly integrate to one between the axes limits shown as a few OM/OC ratio values lie above 2.4, but most of differences are likely due to perception of density curves that vary in both height and width. We have added to the caption of Figure 7 (now Figure 5) where kernel density estimates are first used: "Non-parametric densities are approximated by kernel density estimation (Hastie et al., 2009) in figures."

Reviewer 3

Bürki et al. present a probabilistic modeling framework for estimating organic carbon concentrations and OM/OC ratios from infrared spectroscopy, and they apply it to infrared spectra of PM2.5 samples from 17 monitoring sites of the IMPROVE network. The presented approach is based on previous developments regarding functional group analysis from infrared spectroscopy and statistical calibration strategies for organic aerosol quantification. Here, the authors apply Bayesian calibration to provide plausible estimates for parameters in the probabilistic model, and they obtain OC concentrations and OM/OC ratios consistent with other estimation approaches. The presented work is well within the scope of AMT. However, in its present form the manuscript is difficult to follow especially for atmospheric scientists who do not use linear algebra in their day-to-day work. Thus, the manuscript requires major revisions to improve clarity and to focus on the novelty of the approach before publication in AMT.

We thank the reviewer for the encouraging evaluation.

1. While I appreciate the rigorous explanations in the appendices and in the supplementary material, some parts of the main text either require additional information or should be moved to the supplementary material in order to focus on the main objective of the manuscript. In particular, the introduction of the probabilistic framework in section 1.2 might be revised with the general reader in mind. Also, the description of the cluster analysis might be shortened, and Figures 4 and 6 might be moved to supplement section S3.

   We have substantially revised the Section 1.2 to provide additional explanation for a less specialized audience in mind, and moved the details of the cluster analysis from Section 3.2 to Section S3 of the supplement.

   We had originally considered the exact same possibility of moving Figures 4 and 6 to the supplement prior to initial submission, but had included them for the reason that our interpretation of why the FTIR OM/OC values are similar between 2011 and 2013 are based on the spectral profiles and their cluster membership.

2. When comparing the FTIR estimates of OM/OC with the reconstructed fine mass (RCFM) regression solved both by ordinary least squares (OLS) regression and error-in-variables (EIV) regression, the extensive comparison of OLS and EIV seems distracting. It may be beneficial to briefly introduce both OLS and EIV in section 3.3 but then restrict the comparison in section 4.4 and Figures 8 and 10 to FTIR and only OLS, or only EIV.

   We have relegated comparisons to the Supplemental document when estimates are similar (i.e., MCMC and Laplace for FTIR), but estimates between the two RCFM regression methods are different enough that removal of one from the final presentation could misrepresent the challenges of interpreting OM/OC from the RCFM approach.

   To make the discussion more clear, we have relabeled OLS and EIV as RCFM-OLS and RCFM-EIV.

3. Functionalization by aldehyde, peroxide, aromatic, phenolic, organonitrate, and organosulfate groups is not included in the presented set of calibrations. While it is prudent to prioritize functional groups that are expected to be highly abundant, one has to be very careful when interpreting the results, as stated for example when discussing low mass recovery fractions that "we cannot rule out the need to examine additional FGs" (l.349). For example, organosulfates may become more important in situations when biogenic VOCs are processed in anthropogenically influenced air masses. Thus, changes in OM/OC ratios between 2011 and 2013 observed in the RCFM regression but not observed in the FTIR estimates could also indicate an increasing influence of FGs not taken into account in the presented set of calibrations. I recommend a brief discussion.

   This is a very good point. We have modified/added the following discussion to Section 3.4:

   "This comparison may support the need for further evaluation along two directions. One is in interpreting a_{OC} from RCFM regression as a surrogate for the OM/OC ratio (Hand et al., 2019). The other is in understanding the changing contributions of FGs not included in our set of calibrations (that also are excluded from or have negligible influence on the spectral cluster analysis) over this period. For instance, recent studies of trends in the Southeast US suggest that aromatic, organosulfate, organonitrate, and peroxide-containing compounds in OM have declined in response to reduced anthropogenic emissions of volatile organic compounds, SO_2, and NO_x (the latter two affecting OM through their influence over aqueous-phase reactions and oxidant levels) over the last decades (Pye et al., 2015; Blanchard et al., 2016; Marais et al., 2017; Carlton et al., 2018; Pye et al., 2019). While most of these trends would contradict the direction of discrepancy in OM/OC trends estimated by RCFM and FTIR, the magnitude of changes in emissions and the response of OM likely differ across sites and years considered in this study."

Minor comments:

- Figure 2: What is the reasoning for fitting Weibull distributions to the prior "fractional carbon" coefficients but a normal distribution to the mass recovery fraction?

  The Weibull distribution permits asymmetry, which was important to capture for the two parameters. This has now been noted in the caption.

- Figure 7 is introduced after Figure 4 and before Figures 5 and 6. Please re-order figure numbers.

  We thank the reviewer for catching this error. It has been corrected. Supplement Sections S1 and S2 have now been switched for the same reason.

- Conclusions: The final notion that additional constraints from additional measurements such as NMR or photometry can be added is really helpful and important.

The original manuscript had referred to them in the statement: "or comparison to additional measurements of FGs (Decesari et al., 2007; Ranney and Ziemann, 2016)" but the two methods are now written out explicitly.

Technical comments:

- Please introduce all abbreviations when first used, e.g. "RCFM" in line 141, L-BFGS-B in l.562, etc.

  We thank the reviewer for pointing this out (and all other corrections below) — we have now introduced these abbreviations in the text.

- l.5: "For instance, a subset of model..." - please revise the sentence.

  We have added: "[...] *but* generate substantially different predictions [...]"

- l.14: Carboxylic acid should be "COOH".

  Corrected.

- l.82 "... proposed an extension to this approach..."

  Added: "to"

- l.138/Figure 1: Please indicate the four regions SW/NW/SE/NE by adding lines on the map in Fig. 1.

  This has been done.

- l.165: "L2 norm" might need explanation.

  Now described as "Euclidean distances from the origin when spectra are represented as vectors"

- l.201: Revise "...regression of to y".

  Removed "of"

- l.355: "but it more likely reflects" instead of "but is more likely reflects"

  Corrected.

- l.394: Add parentheses to references.

  Corrected.

- l.451: Remove the extra "from" from "...that estimate OM/OC from from mass balance...".

  Corrected.

- Figure 2: "representing" instead of "representating"

  Corrected.

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.

Anderson, J. A. and Seyfried, W. D.: Determination of Oxygenated and Olefin Compound Types by Infrared Spectroscopy, Analytical Chemistry, 20, 998–1006, https://doi.org/10.1021/ac60023a002, 1948.

Arangio, A., Delval, C., Ruggeri, G., Dudani, N., Yazdani, A., and Takahama, S.: Electrospray Film Deposition for Solvent-Elimination Infrared Spectroscopy, Applied Spectroscopy, 73, 638–652, https://doi.org/10.1177/0003702818821330, 2019.

Blanchard, C. L., Hidy, G. M., Shaw, S., Baumann, K., and Edgerton, E. S.: Effects of emission reductions on organic aerosol in the southeastern United States, Atmos. Chem. Phys., 16, 215–238, https://doi.org/10.5194/acp-16-215-2016, 2016.

Boris, A. J., Takahama, S., Weakley, A. T., Debus, B. M., Fredrickson, C. D., Esparza-Sanchez, M., Burki, C., Reggente, M., Shaw, S. L., Edgerton, E. S., and Dillner, A. M.: Quantifying organic matter and functional groups in particulate matter filter samples from the southeastern United States, part I: Methods, Atmospheric Measurement Techniques Discussions, 2019, 1–39, https://doi.org/10.5194/amt-2019-144, 2019.

Carlton, A. G., de Gouw, J., Jimenez, J. L., Ambrose, J. L., Attwood, A. R., Brown, S., Baker, K. R., Brock, C., Cohen, R. C., Edgerton, S., Farkas, C. M., Farmer, D., Goldstein, A. H., Gratz, L., Guenther, A., Hunt, S., Jaeglé, L., Jaffe, D. A., Mak, J., McClure, C., Nenes, A., Nguyen, T. K., Pierce, J. R., de Sa, S., Selin, N. E., Shah, V., Shaw, S., Shepson, P. B., Song, S., Stutz, J., Surratt, J. D., Turpin, B. J., Warneke, C., Washenfelder, R. A., Wennberg, P. O., and Zhou, X.: Synthesis of the Southeast Atmosphere Studies: Investigating Fundamental Atmospheric Chemistry Questions, Bulletin of the American Meteorological Society, 99, 547–567, https://doi.org/10.1175/BAMS-D-16-0048.1, 2018.

Choy, T. C.: Effective medium theory: principles and applications, Oxford University Press, 2nd edn., https://doi.org/10.1093/acprof:oso/9780198705093.001.0001, 2016.

Coury, C. and Dillner, A. M.: A method to quantify organic functional groups and inorganic compounds in ambient aerosols using attenuated total reflectance FTIR spectroscopy and multivariate chemometric techniques, Atmospheric Environment, 42, 5923–5932, https://doi.org/10.1016/j.atmosenv.2008.03.026, 2008.

Dazzi, A., Deniset-Besseau, A., and Lasch, P.: Minimising contributions from scattering in infrared spectra by means of an integrating sphere, Analyst, 138, 4191–4201, https://doi.org/10.1039/C3AN00381G, 2013.

Debus, B., Takahama, S., Weakley, A. T., Seibert, K., and Dillner, A. M.: Long-Term Strategy for Assessing Carbonaceous Particulate Matter Concentrations from Multiple Fourier Transform Infrared (FT-IR) Instruments: Influence of Spectral Dissimilarities on Multivariate Calibration Performance, Applied Spectroscopy, 73, 271–283, https://doi.org/10.1177/0003702818804574, 2019.

Fischer, K.: Mass absorption indices of various types of natural aerosol particles in the infrared, Applied Optics, 14, 2851–2856, https://doi.org/10.1364/AO.14.002851, 1975.

Griffiths, P. and Haseth, J. A. D.: Fourier Transform Infrared Spectrometry, John Wiley & Sons, In, 2nd edn., 2007.

Hand, J., Prenni, A., Schichtel, B., Malm, W., and Chow, J.: Trends in remote PM2.5 residual mass across the United States: Implications for aerosol mass reconstruction in the IMPROVE network, Atmospheric Environment, 203, 141 – 152, https://doi.org/10.1016/j.atmosenv.2019.01.049, 2019.

Harrick, N. J.: Internal Reflection Spectroscopy, Harrick Scientific Corp., 1979.

Hastie, T., Tibshirani, R., and Friedman, J.: The elements of statistical learning: data mining, inference, and prediction, Springer Verlag, 2009.

Kortüm, G.: Reflectance Spectroscopy: Principles, Methods, Applications, Springer, 1969.

Kuzmiakova, A., Dillner, A. M., and Takahama, S.: An automated baseline correction protocol for infrared spectra of atmospheric aerosols collected on polytetrafluoroethylene (Teflon) filters, Atmospheric Measurement Techniques, 9, 2615–2631, https://doi.org/10.5194/amt-9-2615-2016, 2016.

Leisner, T. and Wagner, R.: Infrared Spectroscopy of Aerosol Particles, chap. 1, pp. 3–24, CRC Press, 2010.

Marais, E. A., Jacob, D. J., Turner, J. R., and Mickley, L. J.: Evidence of 1991–2013 decrease of biogenic secondary organic aerosol in response to SO_2 emission controls, Environmental Research Letters, 12, 054 018, https://doi.org/10.1088/1748-9326/aa69c8, 2017.

Maria, S. F., Russell, L. M., Turpin, B. J., Porcja, R. J., Campos, T. L., Weber, R. J., and Huebert, B. J.: Source signatures of carbon monoxide and organic functional groups in Asian Pacific Regional Aerosol Characterization Experiment (ACE-Asia) submicron aerosol types, Journal of Geophysical Research-Atmospheres, 108, https://doi.org/10.1029/2003JD003703, 2003.

McClenny, W. A., Childers, J. W., Rōhl, R., and Palmer, R. A.: FTIR transmission spectrometry for the nondestructive determination of ammonium and sulfate in ambient aerosols collected on teflon filters, Atmospheric Environment, 19, 1891–1898, https://doi.org/10.1016/0004-6981(85)90014-9, 1985.

Milosevic, M.: Internal Reflection and ATR Spectroscopy, Chemical Analysis: A Series of Monographs on Analytical Chemistry and Its Applications, John Wiley & Sons, Inc., 2012.

Pope, R., Stanley, K. M., Domsky, I., Yip, F., Nohre, L., and Mirabelli, M. C.: The relationship of high PM2.5 days and subsequent asthma-related hospital encounters during the fireplace season in Phoenix, AZ, 2008–2012, Air Quality, Atmosphere & Health, 10, 161–169, https://doi.org/10.1007/s11869-016-0431-2, 2017.

Pye, H. O. T., Luecken, D. J., Xu, L., Boyd, C. M., Ng, N. L., Baker, K. R., Ayres, B. R., Bash, J. O., Baumann, K., Carter, W. P. L., Edgerton, E., Fry, J. L., Hutzell, W. T., Schwede, D. B., and Shepson, P. B.: Modeling the Current and Future Roles of Particulate Organic Nitrates in the Southeastern United States, Environmental Science & Technology, 49, 14 195–14 203, https://doi.org/10.1021/acs.est.5b03738, 2015.

Pye, H. O. T., D'Ambro, E. L., Lee, B. H., Schobesberger, S., Takeuchi, M., Zhao, Y., Lopez-Hilfiker, F., Liu, J., Shilling, J. E., Xing, J., Mathur, R., Middlebrook, A. M., Liao, J., Welti, A., Graus, M., Warneke, C., Gouw, J. A. d., Holloway, J. S., Ryerson, T. B., Pollack, I. B., and Thornton, J. A.: Anthropogenic enhancements to production of highly oxygenated molecules from autoxidation, Proceedings of the National Academy of Sciences, 116, 6641–6646, https://doi.org/10.1073/pnas.1810774116, 2019.

Ramadan, Z., Song, X.-H., and Hopke, P. K.: Identification of Sources of Phoenix Aerosol by Positive Matrix Factorization, Journal of the Air & Waste Management Association, 2011.

Reggente, M., Dillner, A. M., and Takahama, S.: Predicting ambient aerosol thermal-optical reflectance (TOR) measurements from infrared spectra: extending the predictions to different years and different sites, Atmospheric Measurement Techniques, 9, 441–454, https://doi.org/10.5194/amt-9-441-2016, 2016.

Reggente, M., Dillner, A. M., and Takahama, S.: Analysis of functional groups in atmospheric aerosols by infrared spectroscopy: systematic intercomparison of calibration methods for US measurement network samples, Atmospheric Measurement Techniques, 12, 2287–2312, https://doi.org/10.5194/amt-12-2287-2019, 2019.

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

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.

Takahama, S., Dillner, A. M., Weakley, A. T., Reggente, M., Bürki, C., Lbadaoui-Darvas, M., Debus, B., Kuzmiakova, A., and Wexler, A. S.: Atmospheric particulate matter characterization by Fourier transform infrared spectroscopy: a review of statistical calibration strategies for carbonaceous aerosol quantification in US measurement networks, Atmospheric Measurement Techniques, 12, 525–567, https://doi.org/10.5194/amt-12-525-2019, 2019.

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)

Abstract.

The Fourier transform infrared (FTIR) spectra of fine particulate matter ($PM_{2.5}$) contain many important absorption bands relevant for characterizing organic matter (OM) and obtaining organic matter to organic carbon (OM/OC) ratios. However, extracting this information quantitatively — accounting for overlapping absorption bands and relating absorption to molar abundance — ~~poses several challenges. For instance, a subset of model parameters lead to calibrations that test almost indistinguishably well against laboratory standards generate substantially different predictions in ambient samples. Furthermore, additional parameters related to molecular structure are required to estimate carbon content from functional group (FG) abundance. However, since many carbon atoms can be branched (not fully functionalized) or polyfunctional, these parameters are not well constrained for ambient sample mixtures.~~ and furthermore relating abundances of functional groups to that of carbon atoms pose several challenges. In this work, we ~~present a probabilistic framework to characterize combinations of these parameters that are consistent with~~ define a set of parameters that model these relationships and apply a probabilistic framework to identify values consistent with collocated field measurements of ~~organic carbon (OC), for which estimates from~~ thermal optical reflectance ~~(TOR) measurements are used. Uncertainties in this probabilistic framework characterize the plausibility of many different parameter values that yield acceptable predictions (to the extent that they can be evaluated) neglected in conventional estimates of statistical uncertainties. Based on calibrations of aliphatic CH, alcohol COH, carboxylic acid COO, carboxylate COO, and amine NH2, we find model parameters for approximately homogeneous groups of samples determined from~~ organic carbon (TOR OC). Parameter values are characterized for various sample types identified by cluster analysis of ~~FTIR spectra~~ sample FTIR spectra, which are available for 17 sites in the Interagency Monitoring of Protected Visual Environments (IMPROVE) monitoring network (7 sites in 2011 and 10 additional sites in 2013). ~~These groups are interpreted as being predominantly influenced~~ The cluster analysis appears to separate samples according to predominant influence by dust, residential wood burning, wildfire, urban sources, and biogenic aerosols.

~~The resulting calibrations~~ Functional groups calibrations of aliphatic CH, alcohol COH, carboxylic acid COOH, carboxylate COO, and amine NH2 combined together reproduce TOR OC concentrations ~~($R^2 = 0.96$~~with reasonable agreement ($r = 0.96$ for 2474 samples) and provide OM/OC values generally consistent with our current best estimate of ambient OC. The mean OM/OC ratios corresponding to sample types determined from cluster analysis range between 1.4 and 2.0, though ratios for individual samples exhibit a larger range. Trends in OM/OC for sites aggregated by region or year are compared with another regression approach for estimating OM/OC ratios from a mass ~~balance~~ closure equation of the major chemical species contributing to PM fine mass. Differences in OM/OC estimates are observed according to estimation method and are explained through the sample types determined from spectral profiles of the PM.

1 Introduction

Organic mass to organic carbon (OM/OC) was originally characterized using gas chromatograph-mass spectrometry (GC-MS) data (White and Roberts, 1977; Turpin and Lim, 2001) by estimating molecular weight per carbon for individual molecules. However, the analyzed compounds only comprised a small fraction of the overall OM mass and their representativeness for actual aerosol mixtures has been a subject of perennial inquiry. An alternative approach to estimate OM from mass balance of chemical species obtained by sequential extraction has been demonstrated (El-Zanan et al., 2005; Polidori et al., 2008; El-Zanan et al., 2009), but the labor-intensive operation limits the number of samples that can be analyzed. To obtain an effective OM/OC over a large number of samples for a given site or season, regressing concentrations of a suite of particulate matter (PM) components to the gravimetric mass (via the "Reconstructed Fine Mass" equation) in monitoring network measurements has been proposed (Frank, 2006; Malm and Hand, 2007; Simon et al., 2011). However, the results can be difficult to interpret on account of combined measurement errors and intercorrelations among PM component concentrations.

In this work, we advance our ability to estimate OM/OC from Fourier Transform infrared (FTIR) spectra of PM (Allen et al., 1994; Russell, 2003; Takahama and Ruggeri, 2017). In this approach, OM and OC is estimated from organic molecular structures in the PM detected by absorption of mid-infrared radiation. The model for OC estimation from functional groups (FGs), referred to as "FG-OC", and relevant background is presented in Section 1.1. A new framework for constraining estimates through a combination of laboratory and ambient measurements and chemical simulations is described in Section 1.2.

1.1 OM/OC by FG estimation

Another bottom-up approach for deriving estimates of OM/OC is to use chemical measurements of atomic composition of the organic fraction using mass fragments from high resolution aerosol mass spectrometry (Aiken et al., 2008) and FGs from FTIR. Here we focus on FTIR based on its demonstrated capability to characterize $PM_{2.5}$ on Polytetrafluoroethylene (PTFE) filters collected in US monitoring networks. The original concept of calibrating by FGs was outlined by ~~Allen et al. (1994),~~ Anderson and Seyfried (1948) and Allen et al. (1994); and further developed by Russell (2003) and Ruthenburg et al. (2014).

The areal FG-OC mass density m_{C} on each sampled filter i is constructed from the areal molar densities n of several FGs, denoted by the index g:

$$m_{\mathrm{C},i} = \frac{M_{\mathrm{C}}}{\alpha} \sum_{g \in \mathcal{G}^*} \lambda_{\mathrm{C},g} n_{ig} \tag{1}$$

$M_{\mathrm{C}} = 12.01$ is the atomic mass of carbon, α is the mass recovery fraction, and λ_{C} is a coefficient that can be interpreted as the mean "fractional carbon" associated with each FG within the set of measured FGs, $\mathcal{G}^*$. Mass and molar densities typically take on units of $\mu\mathrm{g\,m^{-3}}$ and $\mu\mathrm{mol\,m^{-3}}$, respectively. The molar densities of each FG ~~can be~~ are related to spectral absorbances x by a separate linear model (Ruthenburg et al., 2014):

$$n_{ig} = \sum_{j \in \mathcal{J}} x_{ij} \beta_{jg}^{(k_g)} \ . \tag{2}$$

The ~~basis for~~ approximation made by eq. 2 is that the absorbance due to a substance is proportional to its abundance (Beer-Lambert-Bouguer law) (Griffiths and Haseth, 2007); the coefficients β embody the extent of overlap among target analyte and interferents, and relation between absorbance and molar densities. The coefficients are determined by calibration of laboratory standard spectra to known molar densities of FGs; however, regularization must be used to solve for β because the number of variables (spectral absorbances) are typically greater than number of calibration samples, absorbances are multicollinear, and the inverse solution is sensitive to small perturbations to the data. Partial least squares (PLS) regression (Wold et al., 1983; Martens and Næs, 1991) projects the spectra matrix and areal density of target analyte onto a set of common latent variables, and regularization is imposed by truncating the number of these variables. Therefore, β is a function of the regularization parameter — the number of latent variables k retained — for each FG. Further details for PLS are provided in Appendix B, and a summary of symbols related to the FG-OC model is provided in Table A1.

From the same molar densities of FGs used to estimate m_{C}, molar densities of non-carbon atoms in set $\mathcal{A}^*$ can be added to provide an estimate of OM:

$$(\mathrm{OM})_i = m_{\mathrm{C},i} + \sum_{g \in \mathcal{G}^*} \sum_{a \in \mathcal{A}^*} M_a \lambda_{ag} n_{ig}$$

λ_{ag} are integers relating FG abundances to composition of atoms a, and — unlike $\lambda_{\mathrm{C},g}$ — are well-defined. OM/OC is estimated by normalization to estimated carbon:

$$(\mathrm{OM/OC})_i = 1 + \frac{\sum_{g \in \mathcal{G}^*} \sum_{a \in \mathcal{A}^*} M_a \lambda_{ag} n_{ig}}{m_{\mathrm{C},i}} \tag{3}$$

There are two specific challenges associated with OC estimation from FGs, which also affect OM and OM/OC estimates. The first is to select the appropriate model (β) when a non-unique set of regularization parameters generate similar predictions for laboratory standards used for validation, but vary widely in their predictions in ambient samples (Reggente et al., 2019). The second is to determine a relationship between FG abundance to number of carbon atoms (through λ_{C} and α) since many carbon atoms can be polyfunctional, functionalized with FGs that are not measured, or not functionalized to be detectable by FTIR. The fractional carbon parameter λ_{C} take on values of unity or less to prevent multiplicitous enumeration of the same carbon atom

from knowledge of FG abundance. For instance, $\lambda_{C,aCH} = 0.5$ for methylene carbon leads to the correct estimate of one carbon atom for every two aliphatic CH (aCH) groups measured. Similarly, $\lambda_{C,aCH} = 0.33$ corresponds to methyl carbon, $\lambda_{C,aCH} = 1$ to methine carbon, and so on. Conventionally, λ_C was obtained by assuming the most numerous configurations of carbon present in assumed archetypal molecules (e.g., linear hydrocarbon or ring-structured compounds). Values assumed in previous works range between 0.39 and 0.88 (Allen et al., 1994; Russell, 2003; Reff et al., 2007; Chhabra et al., 2011; Ruthenburg et al., 2014; Table S1); similar uncertainties exist for other FGs. Takahama and Ruggeri (2017) proposed an extension to this approach whereby organic molecules and molecular mixtures are conceptualized as a collection of functionalized carbon atoms. Based on the FGs for which calibrations are built, λ_C can be estimated from the number of measured bonds on each carbon atom or by regression over a collection of carbon atoms. Likewise, the detectable fraction of carbon atoms, α, in molecules and molecular mixtures can be calculated exactly within this scheme. This approach was illustrated for molecules found in the aerosol phase from a simulation of α-pinene photooxidation (in the presence of NO_x) coupled with gas/particle partitioning (Ruggeri et al., 2016).

Parameter selection based on surrogate samples (either laboratory samples or virtual molecules in simulation) and independent estimation of ambient OC and OM is the ultimate objective for operational use of FTIR. However, there are inherent differences in chemical composition (i.e., molecular structure, mixture complexity) between such surrogate samples or mixtures with real, ambient PM. Past studies to evaluate a limited number of parameter selection approaches, however, have led to various degrees of agreement between FG-OC and TOR OC, and it was unclear how this bias was manifested in OM/OC estimates reported by FTIR. Therefore, at the current stage of development, we define our objective to devise a framework to characterize the multitude of plausible parameters that are consistent with available field measurements. Because we do not have reference values for each FG in ambient samples, we turn to available observational data with lower chemical resolution (TOR OC) as reference, together with a probabilistic framework (Section 1.2) for providing plausible estimates for model parameters. Despite known artifacts ~~(Section 2),~~ (Watson et al., 2005; Chow et al., 2005; Cheng et al., 2011; Chan et al., 2019), TOR OC serves as a useful target for FG-OC calibration at this stage to constrain its parameter uncertainties; the implications of these artifacts are also taken into consideration in the model evaluation stage. This ~~procedure for constraining parameter uncertainties in FG-OC leads to~~ strategy furthermore allows estimation of OM/OC from FTIR that are consistent with TOR OC, which is widely used as a reference for OC.

1.2 Probabilistic framework

The inverse problem of parameter estimation in calibration is ill-posed, meaning that small differences in the input — either data or model parameters — may lead to instabilities in the solution (i.e., parameter estimates) (Kabanikhin, 2008; Calvetti and Somersalo, 2018). Bayes' theorem (Bayes, 1763; Robert and Casella, 2010; Gelman et al., 2013) provides ~~the~~ a theoretical foundation for ~~the parameter characterization framework.~~ introducing regularization (i.e., auxiliary knowledge) in natural units of the parameters to stabilize the solution, and for characterizing plausibility of candidate

parameters. Letting p broadly denote any probability density or mass function, the theorem can be written as

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \tag{4}$$

where $p(y) = \int_\theta p(y|\theta)p(\theta)d\theta$. y is the observed data (TOR OC), $\theta = \{\theta_1, \theta_2, \ldots \theta_D\}$ is the parameter vector of dimension D (which includes unfixed FG-OC and PLS parameters), $p(\theta)$ is the prior distribution of parameters, $p(y|\theta)$ is the likelihood, and $p(\theta|y)$ is the posterior distribution. The model for FG-OC (m_C, eq. 1) and explanatory variables (ambient sample spectra, denoted by x in eq. 2) corresponding to each TOR OC observation are assumed given and are excluded in this notation (Gelman et al., 2013). In this multivariate context, a single integral denotes an integral or sum over all parameters. Notation related to probabilistic modeling is summarized in Table A2; data and models used for each of these terms are further described in later sections.

~~The inverse problem of parameter estimation in calibration is ill-posed, meaning that small differences in the input — either data or model parameters — may lead to instabilities in the solution (i. e., parameter estimates) (Kabanikhin, 2008; Calvetti and Somersalo, 2018). The prior~~ As apparent from eq. 4, model parameters are treated as random variables and therefore intrinsically associated with probability distributions. $p(\theta)$ ~~additionally provides a natural~~ serves as the mechanism for regularization~~(incorporating auxiliary knowledge) to constrain the inverse solution; with increasing number of observations effectively overriding~~ , and its influence on ~~the~~ final estimates $p(y|\theta)$ ~~. With θ assumed to be random variables with probability distributions to be characterized according to the agreementbetween model and measurement, parameter uncertainty in this framework reflects the plausibility of different models providing approximately similar predictions~~becomes diminishingly small with increasing number of observations y. $p(y|\theta)$ reflects plausibility of parameters evaluated from model-measurement agreement; the epistemic uncertainty characterized by this distribution (O'Hagan, 2004) is reduced in accordance with informativeness of y. As a point of contrast, conventional ~~model fitting~~ modeling approaches typically rely on expected values of $p(\theta)$ to fix model parameters for forward estimation of y from x; possibly using their distributions for error propagation. The inverse problem is formulated as ~~solving~~ an optimization problem to obtain a point estimate of θ that maximizes $p(y|\theta)$, ~~with parameter uncertainties characterized by confidence intervals and model estimates by prediction intervals in the classical (or frequentist) approach to parameter estimation. Estimates of uncertainty (e.g., confidence intervals) in this case reflect the ability to characterize a single parameter value given the number of samples used and the magnitude of variations present in the data.~~ without incorporating knowledge of $p(\theta)$. Confidence intervals or prediction intervals obtained through this classical approach reflects the aleatoric uncertainty attributed to measurement errors and limitations of statistical sampling (Dowd, 2018).

Bayesian inference has been used previously in atmospheric modeling (e.g., Pinder et al., 2006; San Martini et al., 2006; Thompson et al., 2011; Henderson et al., 2012; Wang et al., 2013; Tukiainen et al., 2016) for estimating under-constrained parameters using field observations in several different contexts. We adopt this approach to provide probabilistic estimates to unknown parameters; starting from prior distributions derived from laboratory measurements and available molecular structures, and updating them based on their plausibility for modeling OC as reported by TOR. In particular, the mass recovery fraction of OC is explicitly included as an unknown parameter for estimation to allow better understanding of potentially

measured and unmeasured contributions of carbon to FG-OC; separate from remaining biases with respect to the TOR measurements. We describe the measurements used in Section 2 and adaptation of this modeling framework in Section 3. Results are presented in Section 4 and concluding remarks provided in Section 5.

2 Experimental data

We apply this method to the Interagency Monitoring of Protected Visual Environments (IMPROVE) (http://vista.cira.colostate.edu/Improve/) 2011 and 2013 data set (2474 samples) used by Reggente et al. (2016) and Takahama et al. (2019), except that the Baengnyeong Island, South Korea, site is excluded to focus on the US sites (Figure 1). The Sac and Fox, KS, site was discontinued mid-2011 and so is not included in the analysis for the 2013 data set. Contiguous US sites are further demarcated into Northeast, Southeast, Southwest, and Northwest regions by the position 40 °N and -100 °W following the convention of Hand et al. (2019). The data set consists of reported values and uncertainties for gravimetry, TOR, X-ray fluorescence (XRF), and ion chromatography (IC), which are used for Bayesian calibration and ~~RCFM regression~~ regression analysis of the reconstructed fine mass (RCFM) equation (Section 3.3). The reported data were obtained from the Federal Land Manager Environmental Database (FED) (http://views.cira.colostate.edu/fed/; last accessed 08/16/2019).

~~For functional group calibration~~

For FG calibrations, we use ~~laboratory standards from Ruthenburg et al. (2014) that includes~~ 250 laboratory standard samples consisting of nine type of organic compounds and organic blanks (ammonium sulfate standards with no organics) previously prepared by Ruthenburg et al. (2014). The calibration models of Kamruzzaman et al. (2018) and Boris et al. (2019) are additionally adapted for quantification of the amine and carboxylate content, respectively. PTFE of ambient and laboratory samples were analyzed nondestructively by FTIR in transmission mode (Maria et al., 2003) after placing them in a custom minichamber purged with air passed through a molecular sieve to remove water vapor and carbon dioxide (Ruthenburg et al., 2014; Debus et al., 2019). Spectra were truncated to the region above 1500 $\mathrm{cm}^{-1}$and baseline corrected (Kuzmiakova et al., 2016) to reduce scattering contributions from the PTFE filter (McClenny et al., 1985) and particles (Takahama et al., 2019). Further details on the sample collection, analysis, and spectra processing steps are described by previous works (Ruthenburg et al., 2014; Reggente et al., 2016; Debus et al., 2019; Takahama et al., 2019). This body of work leads to a collective set of measured functional groups $\mathcal{G}^*$ consisting of aliphatic CH (aCH), alcohol COH (aCOH), carboxylic COOH (COOH), nonacid carbonyl (naCO) (which includes ketone and ester), carboxylate COO (oxOCO), and amine NH2 (NH2). Uncertainties in PLS calibration and molecular structure parameters only associated with aCH, aCOH, and COOH are considered, since the other species did not contribute an appreciable amount to the FG-OC over a range of parameters considered. Because of the inclusion of COOH (for which $\lambda_{\mathrm{C,COOH}} = 1$) and additional fixed contributions from several FGs, the mass recovery parameter α in eq. 1 can be uniquely distinguished from $\lambda_{\mathrm{C,aCH}}$ and $\lambda_{\mathrm{C,aCOH}}$, leading to a model that is identifiable (Walter and Pronzato, 1997).

3 Statistical analysis

3.1 Cluster analysis of spectra

Effective model parameters for a group of samples can be estimated at the level of each site or season directly. However, estimating parameters for a group of chemically similar samples instead is favorable in that parameter values associated with molecular structure are more likely to be representative for each sample in ~~an approximately homogeneous~~ a less diverse population. ~~We use FTIR spectra themselves as an indicator for chemical similarity, and perform cluster analysis to create subgroups interpreted to be chemically similar~~Normalized FTIR spectra are used as indicators of chemical composition and grouped by hierarchical cluster analysis according to similarity (Hastie et al., 2009; Russell et al., 2009) (further details are provided in Section S3). Model parameters are then applied to each member sample and aggregate statistics for OM and OM/OC are obtained for each site and seasons from their constituent samples.~~-~~

~~Hierarchical cluster analysis (Bishop, 2009; Hastie et al., 2009) is used to categorize samples into spectroscopically similar groups (Russell et al., 2009; Liu et al., 2009; Ruthenburg et al., 2014). Spectra are first preprocessed by baseline correction (Kuzmiakova et al., 2016) and wavenumber selection (retaining only regions in the range 3700–2500 and 1820–1500) to reduce the influence of substrate interference, particle scattering, and (carbon dioxide and water) vapors in the analysis chamber (Russell et al., 2009). The spectra are then normalized by their respective L2 norms so that they vary by composition rather than absolute absorbance (which includes the effect of mass loading in addition to composition). Finally, more than 1000 wavenumbers of the normalized spectra matrix are reduced to 9 dimensions using mean-centered, unscaled principal component analysis. These 9 principal components are selected from the eigenvalue profile ("scree plot") and their capability to explain 99% of the variance of the original spectra matrix. While instrumental noise does not contribute much to the overall signal (Debus et al., 2019), this preprocessing step additionally reduces the remaining water vapor contribution to the signal that is visible in spectra with low mass loadings, and makes distance metrics used for characterizing similarity more meaningful than what can be obtained in higher dimensions of correlated variables (Domingos, 2012).~~-~~

~~The Euclidean distance metric with complete linkage is used for clustering samples based on their principal component scores. The number of clusters are heuristically selected by examining how the overall variability is reduced within each cluster (using the within sum-of-squares metric), and how well individual samples are served by the algorithmically-determined associations (with the Silhouette coefficient) with the creation of each additional cluster (Figure S6). Eleven superclusters are selected from this procedure, and model parameters θ estimated for each cluster and applied every member within it to predict FG-OC and FG-OM. As low signal-to-noise ratio samples can adversely affect the operations involving normalized spectra (i.e., principal component and cluster analyses), 10% of samples with the lowest L2 norms are initially excluded in the procedure above, but are assigned to the most appropriate cluster through k-nearest neighbor (k-NN) classification in the principal component space a posteriori for completeness~~.

3.2 Bayesian calibration

The statistical model

$$y_i \sim \mathrm{N}(m_{\mathrm{C},i}, \sigma_i) = m_{\mathrm{C},i} + \varepsilon_i \tag{5}$$

assumes that systematic variations of TOR OC y in each sample i are modeled by FG-OC m_{C}, and non-systematic contributions of measurement errors ε are normally distributed with standard deviation σ (San Martini et al., 2006; Skoog et al., 2017). The likelihood function in this model corresponds to

$$p(y|\theta) = \prod_{i \in \mathcal{S}} \left(\frac{1}{2\pi\sigma_i^2} \right)^{1/2} \exp\left[-\frac{(y_i - m_{\mathrm{C},i})^2}{2\sigma_i^2} \right] \tag{6}$$

where the product is taken over all samples in the set denoted by $\mathcal{S}$.

Choosing a prior distribution $p(\theta)$ (eq. 4) is not a trivial task (Bishop, 2009). Where possible, it is desirable to have an informative but weak prior that does not have disproportionate impact on the results. The prior distribution also imposes bounds on the solution in that the likelihood estimated from eq. 6 are substantially downweighted in near-zero probability density regions specified by the prior (or not considered in regions where the density is identically zero for distributions with finite bounds).

We parameterize the uncertainty σ^2 in eq. 6 as

$$\sigma_i^2 = \sigma_0^2 + \kappa^2 y_i^2 \,, \tag{7}$$

with σ_0^2 denoting the irreducible error and κ^2 denoting a coefficient for the heteroscedastic (concentration-dependent) error. These terms have familiar interpretations as $2\sigma_0$ is a typical measure of the minimum detection limit (MDL), and κ corresponds to the relative standard deviation (σ/y) in the limit of high concentrations ($y \gg \sigma_0$). σ^2 for each sample is calculated from combined uncertainties of the thermal fractions of TOR OC, and initial estimates for these two parameters are obtained via regression ~~of~~ to y. As reported to the IMPROVE database, TOR OC uncertainties are assumed independent across samples, and correlation of errors across thermal fractions for each sample are omitted. σ_0 is kept fixed to the fitted value of 0.04 $\mu\mathrm{g\,m^{-3}}$ as $2\sigma_0$ is higher than that reported for the TOR OC MDL (0.05 $\mu\mathrm{g\,m^{-3}}$) (Dillner and Takahama, 2015) and serves as a conservative estimate. The fitted κ is approximately 7%, which is lower than collocated precision or overall errors ($\kappa \sim 15\%$) that have been reported elsewhere (Dillner and Takahama, 2015; Brown et al., 2017). Therefore, we include κ^2 as an additional unknown parameter to be estimated, and assume a inverse gamma distribution around the fitted value for the prior. Uncertainties in n and molecular structure parameters due to model variance of eq. 2 and C2 are not included in this estimate. The analytical precision (typically within 5%) is greater than that of TOR (Debus et al., 2019) but collocated precision can be similar in magnitude (Dillner and Takahama, 2015). Incorporating these considerations into eq. 6 poses additional challenges (Rock et al., 1977) and are not considered for this study. Because of the heteroscedastic error model (eq. 7), samples with lower concentrations can have comparable or greater impact on the likelihood; the abundance of lower concentration samples (according to approximately lognormal concentrations in atmospheric samples; Ott, 1994) means a few high concentration points have less influence on parameter estimation (Section S2).

To estimate probabilities associated with the number of PLS latent variables, We use mean square error of cross validation (MSECV) typically used for model selection and convert them into probabilities using Boltzmann weighting (Appendix C1). The proposed approach leads to a prior favoring solutions with lower MSECV estimated for the calibration set (laboratory standards) and downweighting substantially high-bias (high MSECV) solutions which are not sufficiently complex to capture the spectral variations for quantification of the FG (Figure S1).

The priors for structural parameters $\lambda_{C,g}$ and α are estimated from virtual mixtures of primary organic aerosol compounds from automobile exhaust and wood burning measured by GC-MS (Rogge et al., 1993, 1998), and secondary organic aerosol compounds in the Master Chemical Mechanism v3.3.1 database (Jenkin et al., 1997; Saunders et al., 2003). In both data sets, compounds likely to be in the aerosol phase were selected based on volatility (equilibrium vapor concentration $C^0 \leq 10^{3.5}$ $\mu g\,m^{-3}$) (e.g., Robinson et al., 2007). Further details of the method are provided in Appendix C2 and results of analysis in Section 4.1.

Having defined the likelihood function and prior distributions, we obtain the posterior probability $p(\theta|y)$ from measurements y in two ways. Our primary technique is Markov Chain Monte Carlo (MCMC), which evaluates the unnormalized posterior $p(y|\theta)p(\theta)$ for numerically sampled values of θ. We also confirm our results using Laplace's method, which is a Gaussian approximation about the maximum of the unnormalized posterior. This method can only be used for continuous variables, so it is applied for each combination of k_g. More details on these techniques are provided in Appendix D.

3.3 Reconstructed fine mass regression

For comparison, we estimate OM/OC as interpreted by coefficients of the ~~reconstructed fine mass (RCFM) equation used by IMPROVE~~ RCFM equation (a statement of mass closure) used by the IMPROVE network (Malm et al., 1994; Malm and Hand, 2007; Chow et al., 2015). Given the atmospheric concentration ($\mu g\,m^{-3}$) c of a substance, regression is used to obtain coefficients a:

$$c_{FM} - c_{EC} - c_{SS} = a_{AS}c_{AS} + a_{AN}c_{AN} + a_{OC}c_{OC} + a_{dust}c_{dust} \tag{8}$$

FM is the dry fine mass concentration, measured by gravimetric analysis and corrected for particle bound water using available relative humidity measurements of the analysis laboratory and hygroscopic growth factors for constituent species as described by Hand et al. (2019). *AS* and *AN* are ammonium sulfate and nitrate, respectively, estimated from the sulfate and nitrate under the assumption of full neutralization. *SS* is sea salt, estimated as 1.8 times the chloride concentration. *dust*, also referred to as "soil," is calculated from assumed oxide forms of silicon, calcium, iron, and titanium. *OC* and *EC* are as quantified by the TOR method (Section 2). To reduce collinearity among variables, EC and SS are not included in the regression but subtracted from FM a priori (Simon et al., 2011; Hand et al., 2019). The coefficients and their confidence intervals are obtained by MLR solved by ordinary least squares (OLS) (Weisberg, 2013) and Error-in-Variables regression (EIV) (Fuller, 1987) as described by Hand et al. (2019) and Simon et al. (2011), respectively. To avoid confusion with other approaches described in this study, ~~OLS and EIV regression~~ the two methods for solving eq. 8 will be collectively referred to as RCFM regression and labeled as RCFM-OLS and RCFM-EIV. Furthermore, the results of a_{OC} will be referred to as the OM/OC ratio estimate according to this

approach. OLS does not consider heteroscedasticity or relative magnitude of measurement errors of any variable, which can lead to biased coefficient estimates and confidence intervals that do not reflect the actual uncertainty (Fuller, 1987; Simon et al., 2011; Weisberg, 2013). The latter issue is addressed in this work by providing confidence intervals obtained by bootstrapping (Davison and Hinkley, 1997). EIV regression alleviates this problem by considering measurement errors of both explanatory and response variables explicitly (neglecting error covariances in this implementation); however, the estimates are subject to the accuracy of estimated measurement errors. The implementation provided by Simon et al. (2011) is used for estimation of coefficients and their uncertainties. Analytical uncertainties reported for each measurement are used for their estimates, but unaccounted systematic biases can affect the coefficient a_{OC} (Hand et al., 2019).

4 Results

For this paper, we limit our focus on topics related to the estimation of parameters that generate FG-OC congruent with TOR OC concentrations, and comparisons of new OM/OC ratios obtained by FTIR with RCFM regression estimates. Obtaining FG composition for each filter sample enables analysis of site-specific OM/OC ratios and source-class characteristics in much greater detail, and is reserved for a separate, dedicated paper on the subject. The following subsections cover characterization of prior distributions estimated for the unknown molecular structure parameters λ_{C} and α (Section 4.1), description of spectral clusters formed (Section 4.2), posterior parameter estimates (Section 4.3), and comparison with RCFM regression (Section 4.4)

4.1 Prior distributions

Prior distributions of structural parameters obtained by the method described in Section 3.2 are summarized in Figure 2. The values between 0.46–0.48 for $\lambda_{\mathrm{C,aCH}}$ are consistent methylene ($\mathrm{CH2}$) group structures, though another reason this narrow distribution can occur is that single aliphatic CH bonds are often found together with one other measured FG on the same carbon atom (Takahama and Ruggeri, 2017). In such cases, a value of $\lambda_{\mathrm{C,aCH}}$ close to 0.5 prevents double counting of carbon by the two bonds (Section 1.1). The broad values for $\lambda_{\mathrm{C,aCOH}}$ reflect the diverse carbon types in which alcohol groups are found. The α value centered around 0.74 reflects the undetected carbon fraction, typically missed due to branched molecular structure or functionalization by unmeasured FGs.

Several examples for molecules with incomplete carbon recovery ($\alpha < 1$) are shown in Figure 3. More generally, the types of carbon atoms undetected vary widely in their structure (Figure S3). These molecules contain unfunctionalized carbon atoms (only bonded to other carbon atoms) and carbon atoms functionalized by, for example, aldehyde, peroxide, aromatic, phenolic, and organonitrate groups, which have absorption bands in the mid-infrared but are not included in our set of calibrations. These FGs have not been prioritized for calibration following the hypothesis that molecules with these functionalities are not found in great abundance in IMPROVE samples. Aldehydes are susceptible to hydration in aqueous solutions, leading to formation of alcohols (Schwarzenbach et al., 2002; Takahama et al., 2013b). Peroxides have been shown to be labile under various (light and dark) conditions (Epstein et al., 2014; Krapf et al., 2016). Phenolic OH and aromatic groups exhibit sharp absorption

peaks near 3500 and 3100 cm^{-1}, respectively (Bahadur et al., 2010), which are not observed in ambient sample spectra; in previous studies, Russell et al. (2011) suggested the aromatic and unsaturated FGs contributed to less than 5% of OM mass. Organonitrates also hydrolize in the presence of water to form alcohols and nitric acid (Liu et al., 2012; Zare et al., 2019), and organosulfate FGs are not included in this analysis but their contribution to the overall OM mass concentration is often bound to be less than a few percent (Hawkins et al., 2010; Russell et al., 2011; Takahama et al., 2013a; Budisulistiorini et al., 2015; Hettiyadura et al., 2017). Additionally, oxygen has been found to be the heteroatom contributing most to the variability OM/OC ratios in ambient samples (Pang et al., 2006).

The procedure of parameter updating with ambient OC estimates can help place these values in the proper context. Previous estimates of FG-OM generally reported agreement of 70–100% for submicron OM compared against AMS (Russell et al., 2009; Gilardoni et al., 2009; Corrigan et al., 2013), and FG-OC was estimated to be 60–70% of TOR OC in $PM_{2.5}$ in the IMPROVE network samples (2011 data set) (Ruthenburg et al., 2014; Reggente et al., 2019). While these differences have been partially attributed to incomplete mass recovery of carbon by FTIR, now the estimated mass recovery fraction based on molecular structure information is included explicitly into the calibration model.

In reporting OM/OC using eq. 3, we can expect a systematic underestimation of OM/OC on account of unmeasured FGs. An alternative estimate can be obtained by considering the OM/OC of only the measured, functionalized carbon (i.e., using αm_C for normalization in eq. 3). This latter approach can on average lead to a more representative value of the overall OM/OC (Figure S4) in oxygenated aerosol. For this work, we use eq. 3 which likely provides a lower bound on OM/OC and a means to gauge improvement in OM/OC estimates with the inclusion of additional FG calibrations.

4.2 Cluster descriptions

While the primary objective of cluster analysis for this study is to create chemically ~~homogeneous~~ similar groups for parameter estimation, we include a brief remark on interpreted source classes or composition associated with each spectra type. For this analysis, we use spectral characteristics visualized in Figure 4, concentrations of tracer species or magnitude of tracer variables (Figure S7; consisting of RCFM components and additional trace elements analyzed by XRF), and location and time of occurrence as indicators of source classes (Figure S8).

Clusters 1 and 4 are high sulfate, low organic samples found predominantly in rural areas; suggesting the likely association of the organic fraction with biogenic secondary organic aerosol (SOA); samples in cluster 1 are found predominantly in the SE and NE with a notable absence in the Southwest. Nearly half of samples in Clusters 2 and 5 are found in urban areas — particularly in Phoenix, AZ — and the remaining found in rural areas are likely influenced by nearby urban sources. Clusters 3, 8, and 11 occur predominantly in the Southwest and are associated with mineral dust as evidenced by sharp Si-O-H peaks above 3500 (Reggente et al., 2019), and supported by observations of elevated contributions of elements: Al, Ca, Fe, Si, and Ti. Clusters 6 and 7 occur predominantly in the Southeast and largely consist of samples originally identified by Ruthenburg et al. (2014) as being “anomalous” in their agreement of FG-OC with respect to TOR OC. Reggente et al. (2019) later proposed that these samples contained large ammonium sulfate and ammonium nitrate particles (consistent with IC concentrations) that exhibited an optical artifact known as the Christiansen peak effect, which leads to an increase in transmittance in the

vicinity of the wavelength where i) the refractive index of the substance approaches that of air and ii) the particle size and wavelength of radiation also become similar ($\sim$3300 cm^{-1}). Thus, these samples share a particular absorbance profile and quantification based on assumption of Beer-Lambert law can be challenged in some wavenumber regions — especially near the absorption band of alcohol aCOH — for these samples. Samples in clusters 9 and 10 are associated with burning. For purposes of interpretation, cluster 9 is split according to child nodes of the hierarchical clustering tree into wildfire (cluster 9a) and residential wood burning (cluster 9b) groups, which are labeled according to their occurrence during a known fire period (Rim) and during winter months where residential burning takes place (Phoenix, AZ; Ramadan et al., 2011; Pope et al., 2017) (more in Section 4.3).

Previous work in cluster analysis with aerosol FTIR spectra resolved differences among urban (fossil fuel combustion), terrestrial vegetation (burning and non-burning), and marine aerosols (e.g., Russell et al., 2009; Liu et al., 2009; Takahama et al., 2011; Corrigan et al., 2013). These studies focused on spectra collected during short, intensive field campaigns (typically considering samples from a single location and single season each) with higher time resolution (typically four hours), and used an inlet with nominal size cut of one micrometer. Spectra types from monitoring networks are not expected to have a direct correspondence to their work due to the use of a 2.5 micrometer size cut (more influence of dust and larger inorganic particles) and time resolution (24 hours) of measurements (more mixing of source classes and degrees of aging). In particular the naCO fraction in IMPROVE network samples have been estimated to be negligible using several methods (Reggente et al., 2019), while naCO varies substantially across spectra types in the submicron samples collected during intensive field campaigns and have been used as indicators of biogenic and biomass burning aerosol (Russell et al., 2011). Nevertheless, some similar spectra categories are found through differences apparent in absorption profiles.

That such a large number of samples from a wide range of sites and seasons are considered together in this work suggests that selecting a limited number of clusters for statistical estimation is likely to provide only a crude separation in chemical and spectral variations that differentiate source classes or mixture proportions of source classes. In addition, first differentiation in spectra (i.e., initial branches of the hierarchical tree) is determined by ammonium NH, alcohol aCOH, and carboxylic COH, as their broad absorption bands comprise a substantial portion of the absorbance in the spectrum. These factors can lead to clusters which contain both rural and urban samples that differ primarily by aliphatic CH absorption (which affects the overall OM/OC but not its oxygenated fractionation), and surprising associations across regions (e.g., Fresno, CA, samples associated with samples in the SE in the same cluster). However, for the purposes of parameter estimation this level of disaggregation is found to be computationally tractable and sufficient in that estimates for smaller subsets of spectra do not substantially change the OM/OC estimated with this limited number of clusters.

4.3 Estimated parameters

Estimates of parameter distributions obtained by MCMC are generally confirmed by the Laplace method (Figure 5 shown as an example for a single cluster and Figure S9 for all clusters). Therefore, the following results will focus on results of MCMC analysis. The posterior distributions for most parameters show a departure from the mode of their prior distributions, suggesting that the results are not dominated by influence of the priors. The mode of each posterior parameter distribution for every cluster

is shown in Table 1. The number of latent variables k_{aCH} and k_{aCOH} vary by cluster, suggesting that a different model is appropriate for different spectra types (and presumably different types of PM). The mass recovery fraction α ranges between 0.57 and 0.83 consistent with the range estimated for primary and secondary OM species (Section 3.2). Given our expectations for low abundance of unmeasured FGs (Section 4.1), low α may indicate a surprising amount of branched molecules with unfunctionalized carbon atoms — though we cannot rule out the need to examine additional FGs or that some systematic discrepancies (e.g., in absorption coefficients) between molecules in laboratory and ambient samples are also incorporated into parameter estimates. λ_{aCH} is consistently near 0.48, at the exception of cluster 3, but possibly due to the strong prior. λ_{aCOH} varies much more substantially across clusters and this is likely due to the different configurations of the carbon atom functionalized by aCOH. The coefficient κ for heteroscedastic measurement error varies between 0.13 and 0.31, which is greater than the reported TOR OC analytical error of 0.07. The variations in κ across clusters may partially reflect differences in thermal fractions or sensitivity to different types of compounds, but ~~is~~ it more likely reflects the range of discrepancies between modeled and measured OC across samples that arises from a given set of parameter values. Nonetheless, the estimates of remaining parameters are robust with respect to this assumption, as assessed with simulations in which κ is kept fixed at the prior estimate of 0.07.

The comparison of fitted FG-OC with reference TOR OC (Figure 6) with 95% intervals of the posterior predictive distribution (Robert, 2007; Vehtari and Ojanen, 2012; Gelman et al., 2013; Section S4) shows reasonable agreement ~~($R^2 = 0.96$)~~with regards to correlation and bias. There is an underprediction for several high concentration samples due to the larger number of samples with lower concentrations that collectively influence the likelihood (Section S2). Posterior predictive distributions are symmetric, and FG-OC estimated from their modes are almost identical to that obtained from single-point estimates of parameters obtained as the mode of their respective distributions (Figure S11). TOR OC measurements are out-of-range of 95% prediction intervals of the posterior distribution approximately 5% of samples. No abnormalities are detected in spectra upon investigation, which may indicate that these samples are not well-served by the current calibration model (e.g., the selection of calibration standards). That the cluster containing anomalous samples (clusters 6 and 7) can reproduce TOR OC — in contrast to previous works of Ruthenburg et al. (2014) and Reggente et al. (2019) — is surprising, but that the alcohol aCOH is estimated to be zero can be due to the effect of anomalous dispersion (Section 4.2) and some compensation may be incorporated into the value of α for these samples.

Figure 7 shows the mean OM and OM/OC for each spectra type. Trends in OM estimates across these types are consistent with trends in TOR OC, with burning samples (clusters 9 and 10) exhibiting the highest OM and biogenic and dust-related samples (clusters 1, 3, 4, 8, and 11) having the lowest OM, on average. Samples with urban influences (clusters 2 and 5) have, on average, lower OM/OC than those more associated with oxidized, biogenic (clusters 1 and 4). The high alcohol aCOH contribution to OM/OC in the dust samples (clusters 3, 8, and 11) may be indicative of condensed secondary OM (Murphy et al., 2006; Hawkins et al., 2010; Takahama et al., 2010), but may also partially be due to misappropriated hydroxyl groups or hydrates of water associated with inorganic substances (Hudson et al., 2008; Frossard and Russell, 2012). Wildfire burning samples (cluster 9a) consistently display higher OM/OC than residential wood burning samples (cluster 9b). Because these two

sample types occur during warm and cold months, respectively, the contribution of photochemical aging relative to emission characteristics cannot be easily determined from this type of analysis.

Some variability in OM/OC across samples are present within several clusters. For instance, cluster 9 of the eleven original clusters exhibited a bimodal distribution in OM/OC from distinguishable contributions from urban wood burning and rural wildfire samples (Figure S10, and have already been disaggregated for discussion (Section 4.2). Within clusters 1, 2, and 5, contrast in OM/OC ratios between samples from urban and rural sites can be observed, with values lower by ∼0.2 in the former. Further inspection of child nodes do not clearly separate urban and rural samples as with cluster 9, and this is largely because urban and rural samples in the same cluster differ primarily by the aliphatic aCH content while the oxygenated groups are present in similar proportions. Due to its sharp peaks, aCH absorbances comprises a small portion of the overall variation in spectra considered in the clustering technique and does not exhibit substantial influence in cluster determination. The OM/OC distribution samples in clusters containing dust-influenced samples are broad (regardless of site type) due to the high variability in estimated alcohol aCOH content.

4.4 Spatial and temporal characteristics

A large number of samples are required to evaluate meaningful difference in coefficients due to the number of RCFM components, range of variations in their concentrations, and their combined measurement errors. Therefore, multiple sites or multiple years of data for a given site are often used for analysis ~~Simon et al. (2011); Hand et al. (2019)~~ (Simon et al., 2011; Hand et al., 2019). For this work, we report coefficients for the combined years of 2011 and 2013 and sites aggregated by region (restricted to those for which FTIR spectra are available, Section 2) to examine spatial and seasonal differences, or six sites for which FTIR spectra are available both years to examine temporal trends between the two years.

Estimates across regions and seasons for the two years combined are shown in Figure 8. Given the limited number of sites analyzed in this work, the region labels are used only to summarize results across multiple sites and may not be indicative of results for the entire region. For instance, the highest OM/OC estimated by ~~OLS~~ RCFM-OLS for all (∼160) IMPROVE sites between 2011 and 2015 were found in the Southeast and Northeast regions (Hand et al., 2019), whereas their annual average values are, on average, below that of the Northwest region according to the sites and years considered in this study.

Estimated trends in OM/OC between ~~EIV and OLS~~ the two RCFM regressions are consistent in that they generally predict higher OM/OC during spring and summer, except in the Northwest sites where the highest OM/OC is observed in the winter. This type of agreement is not unexpected as the two methods use the same mass balance approach and concentration measurements. However, OM/OC estimates from ~~OLS~~ RCFM-OLS (ranging between 1.4–2.5) generally underestimates that from ~~EIV~~ RCFM-EIV (1.5–3.1) by ∼0.3 on average. This pattern of underestimation was also reported previously (Simon et al., 2011) — this difference may be partly due to the disproportionate impact of high OC (and low OM/OC) samples on squared residuals and subsequent regression coefficient estimates by ~~OLS~~RCFM-OLS, which are downweighted by uncertainties in ~~EIV~~ RCFM-EIV that increase together with concentration. The large confidence intervals for the Northwest and Northeast sites reflect the fact that only one or two sites are included in these regions, and displays the limit of resolution by the RCFM regression approach for limited sample sizes. Smaller confidence intervals shown for FTIR estimates reflect the fact that re-

gional estimates are calculated as the mean of OM/OC values obtained for each sample. Magnitude of uncertainties in FTIR OM/OC due to posterior parameter uncertainties (Hoff, 2009) for any individual sample is typically below 6%, but can be higher for samples in two clusters (Section S4).

FTIR estimates of OM/OC for these regions (1.7–2.2) are on average more similar to ~~OLS than EIV~~ RCFM-OLS than RCFM-EIV but show less variability across regions and seasons. In general, we expect that FTIR estimates reported here may be conservative (low) if important FGs are missing in our calibration models (Section 4.1). While mean OM/OC ratios and FG composition can be estimated for each location or period explicitly, its magnitude can be roughly anticipated by i) the frequency of cluster types (Figure S13) and ii) variability of OM/OC within each cluster (i.e., urban samples having lower OM/OC in each cluster; Section 4.3). Disaggregating FTIR estimates by site type reveals that seasonal differences are greater in urban areas (∼0.2 between winter and summer) while less pronounced in rural areas (Figure 9); regional averages are more indicative of trends in the latter because there are fewer urban sites and hence smaller number of samples. OM/OC distributions indicate that rural samples over all seasons and urban samples during the summer have a mode close to 1.8, which is the assumed OM/OC multiplier currently assumed for the IMPROVE network. Phoenix, AZ, is an urban site that exhibits particularly extreme differences in OM/OC, with low values due to wood burning and possibly less aged urban emissions in the winter (cluster 9b and 5, respectively), and high values from the influence of dust particles in the spring and summer (clusters 5 and 8) (Figure S13). The broad OM/OC distribution during these warmer months is due to the variability in alcohol aCOH contribution estimated for the dust-influenced samples. More generally, the higher OM/OC ratios estimated for the Southwest sites — particularly HOOV (Hoover, CA), BLIS (Bliss, CA), and MEVE (Mesa Verde, CO) — during the spring season are due to the prevalence of dust-impacted samples. Because organic mass loadings of these dust-impacted samples are relatively low (Section 4.2), the mean OM/OC values during spring are similar to that of summer months if ratios are alternatively calculated taking OC-weighting into account. The higher OM/OC estimated during the spring (1.93) in comparison to summer (1.76) in the single Northeast site (Proctor Maple, VT) is not confirmed by the other two methods as their seasonal differences are not statistically significant, but inspection of spectra types indicates that the biogenic-type samples (cluster 4) were prevalent during the spring while more urban-influenced samples (cluster 5) with lower OM/OC values were found during the summer in comparison.

Considering only the six sites — Phoenix, AZ; Olympic, WA; Proctor Maple, VT; St. Marks, FL; Mesa Verde, CO; and Trapper Creek, AL — for which FTIR measurements are available between 2011 and 2013, we compare differences in mean OM/OC ratios (Figure 10). Hand et al. (2019) previously reported increasing trends in mean OM/OC ratios between 2011 and 2013 over the entire network; particularly with an increase of ∼0.2 during summer months. ~~OLS and EIV~~ RCFM-OLS and RCFM-EIV for these sites also show increasing OM/OC (by 0.35 and 0.5, respectively) for the summer months for the subset of sites analyzed in this work, and a difference of 0.4 is also significant for ~~OLS~~ RCFM-OLS for the spring months. However, FTIR estimates show no such trend, and the FG composition is also remarkably consistent between the two years at these sites (Figure 11). The sample type composition determined by the FTIR spectra between the two years are also similar (Figure S14), which explains this similar estimate of OM/OC. Inspection of other regression coefficients of eq. 8 indicate other changes such as decrease in a_{dust} between the two years, which may suggest changing atmospheric composition

or changes in analytical bias (Hand et al., 2019) that affect estimates of a_{OC}. This comparison ~~reinforces~~ may support the need for further evaluation ~~to interpret~~ along two directions. One is in interpreting a_{OC} from RCFM regression as a surrogate for the OM/OC ratio (Hand et al., 2019). The other is in understanding the changing contributions of FGs not included in our set of calibrations (that also are excluded from or have negligible influence on the spectral cluster analysis) over this period. For instance, recent studies of trends in the Southeast US suggest that aromatic, organosulfate, organonitrate, and peroxide-containing compounds in OM have declined in response to reduced anthropogenic emissions of volatile organic compounds, SO_2, and NO_x (the latter two affecting OM through their influence over aqueous-phase reactions and oxidant levels) over the last decades (Pye et al., 2015; Blanchard et al., 2016; Marais et al., 2017; Carlton et al., 2018; Pye et al., 2019). While most of these trends would contradict the direction of discrepancy in OM/OC trends estimated by RCFM and FTIR, the magnitude of changes in emissions and the response of OM likely differ across sites and years considered in this study.

5 Conclusions

We presented a new framework to enable estimation of OM and OM/OC from FG calibrations of FTIR spectra that are also consistent with the current best estimate of ambient OC, which is taken from TOR measurements. In contrast to RCFM regression approaches that estimate OM/OC from ~~from~~ mass balance of all other major components contributing to particulate fine mass, estimation of this metric by FTIR uses spectra of particles collected on PTFE filters together with laboratory standards of organic molecules. In contrast to standard multivariable optimization approaches for parameter estimation, the proposed probabilistic approach incorporates prior knowledge of model parameters based on performance against laboratory standards and sensible structural parameter values derived from atmospherically-relevant molecules compiled from measurements or computer models. While this information was exclusively used for parameter determination in previous works, the Bayesian framework used here weighs plausibility of parameter values against ambient observations. The clustering approach used for selecting subgroups with similar spectral profiles also leads to estimation of model parameters that better reflect samples in each subgroup, and provides a way for associating model parameters and OM/OC estimates to various chemical classes of PM.

Model parameters that reproduce TOR OC measurements could be found for more than 94% of samples; this approach also identifies samples for which calibration models are potentially unsuitable. Spectra types associated with dust, wildfire, residential wood burning, urban, and biogenic-influenced samples were found in the IMPROVE 2011 and 2013 samples. Mean OM/OC ratios for various locations or periods are consistent with occurrences of these spectra types. In contrast to RCFM regression methods, no consistent increase in OM/OC was found between 2011 and 2013, and the spectra type composition was also consistent between the two years.

This work enables many directions for future studies. OM/OC ratios and FG composition can be further related to sources and specific sites or seasons for the samples introduced in this calibration study. Furthermore, the framework is described generally such that it can be applied to samples in monitoring networks or chamber experiments, and systematically evaluate improvements in calibrations with new standards or FGs. Parameters that can be applied to new samples for prediction can potentially be determined by assessing spectral similarity of new samples to the sample types es-

tablished through cluster analysis. For increasingly refined spectral types, hierarchical Bayesian modeling (Gelman and Hill, 2007) can be used to model relationships among subgroups (e.g., spectral clusters) overcome limitations in dealing with smaller sample sizes, albeit with added complexity. Additional constraints — such as residual FM (Boris et al., 2019) or ~~comparison to~~ additional measurements of FGs ~~(Decesari et al., 2007; Ranney and Ziemann, 2016)~~ by NMR or spectrophotometry (Decesari et al., 2007; Ranney and Ziemann, 2016; Duarte and Duarte, 2017) — can be introduced to the maximum likelihood expression to explore solutions which are consistent with other available measurements.

Appendix A: Notation

Table A1 describes mathematical symbols for carbon estimation model and Table A2 for Bayesian modeling.

Appendix B: Partial least squares calibration

The origin of the regularization term in eq. 2 specifically for PLS regression is explained in this section. The nonlinear iterative least squares (NIPALS) algorithm (Wold et al., 1983) is used to project a matrix of mean-centered laboratory standard spectra with absorption x_{ij}, defined for each wavenumber j (indexed from 1 to J) and sample i, onto a basis set of spectral profiles (loadings) whose elements are $p_{\ell j}$, with ℓ representing the index of the reduced dimension (also referred to as a latent variable or component). The PLS scores $t_{i\ell}$ embody both the contribution of component ℓ to the spectra and its contribution to the FG abundance (determined by gravimetric analysis for known aerosol composition) after additional scaling by coefficient $q_{\ell g}$:

$$\begin{aligned} n_{ig}(k_g) &= \sum_{j=1}^{J} x_{ij}\beta_{jg}^{(k_g)} + e_{ig} = \sum_{\ell=1}^{k_g} t_{i\ell}q_{\ell g} + e_{ig} \\ x_{ij}(k_g) &= \sum_{\ell=1}^{k_g} t_{i\ell}p_{\ell j} + e_{x,ij} \end{aligned} \qquad \forall\, g \in \mathcal{G}^* \tag{B1}$$

For a selected value of k_g, the components beyond k_g+1 comprise the residual terms $e_{x,ij}$ and e_{ig}. Using the provided training samples, q, and p are found such that the new variables t maximize the covariance with n during the calibration process. Each new spectrum (of laboratory and ambient samples) are then projected onto this basis set and its scores used to estimate the FG abundance.

Appendix C: Estimation of priors

C1 Number of latent variables *k*

For each FG, we estimate a prior for the number of latent variables (denoted as k rather than k_g in this section for readability) by Boltzmann weighting (Adamson, 1979) of their mean squared error of cross validation (MSECV) from laboratory calibrations.

The MSECV is written in terms of the chi-square statistic χ^2:

$$p(k) = \frac{\exp\left(-\chi_k^2/2\right)}{\sum_{k=1}^{K}\exp\left(-\chi_k^2/2\right)} \text{ where } \chi_k^2 = \frac{N \cdot \mathit{MSECV}_k}{s^2} \,. \tag{C1}$$

s^2 is the expected magnitude of error, which we use as a scaling variable fixed to the condition that $\chi^2/(N-k-1) = 1$ (reduced chi-square is unity) for the minimum MSECV solution. The form of eq. C1 is also consistent with the notion of likelihood ratios used in model selection and Akaike weighting (Burnham and Anderson, 2003). The upper limit on k is selected to balance inclusiveness of plausible solutions against computational considerations; for each component k is chosen to include several solutions within one standard error of the MSECV and exclude physically unrealistic ones (with high proportion of negative predictions in concentration). The choice of upper limit for k can change the overall probability, but the relative probability among solutions remain approximately similar for a range of upper limits considered.

C2 Carbon fractions λ_{C} and mass recovery fraction α

This work extends the approach of Takahama and Ruggeri (2017) to study functionalization at the level of each carbon atom for a larger set of atmospherically-relevant molecules with known structure. We consider the set of molecules in primary aerosols $\mathcal{M}_{\text{primary}}$ from GC-MS measurements by Rogge and co-workers (Rogge et al., 1993, 1998) previously analyzed for FG composition by Ruggeri and Takahama (2016); and the set of gas-phase photooxidation products $\mathcal{M}_{\text{secondary}}$ from MCM v3.3.1. Considering species with equilibrium vapor concentrations $C^0 \leq 10^{3.5}$ $\mu\mathrm{g\,m^3}$, there are 193 molecules in $\mathcal{M}_{\text{primary}}$ and 1221 molecules in $\mathcal{M}_{\text{secondary}}$ (Figure S2).

A subset of molecules $\mathcal{M}^{(s)}$ are constructed by varying the fraction ζ of primary vs. secondary aerosol molecules between 0 and 1 by 0.05 increments, and randomly sampling from the required number from each population to satisfy the balance:

$$|\mathcal{M}^{(s)}| = \zeta^{(s)}|\mathcal{M}^{(s)}_{\text{primary}}| + (1-\zeta^{(s)})|\mathcal{M}^{(s)}_{\text{secondary}}|$$

where $|\cdot|$ denotes the cardinality (number of elements) of the set. To accommodate the limited number of primary compounds available for random selection, the total number of molecules $|\mathcal{M}^{(s)}|$ considered for any subset was 50–150 so that each contained a random subset of $\mathcal{M}_{\text{primary}}$ even for $\zeta^{(s)} = 1$. We therefore estimate λ_{C} by nonnegative least squares regression of measurable carbon abundance on FG abundances repeated over various subsets s:

$$n^*_{\mathrm{C},i} = \sum_{g\in\mathcal{G}^*} \lambda^{(s)}_{\mathrm{C},g}\, n_{ig} + e_i \quad \text{where} \quad n^*_{\mathrm{C},i} = \sum_{k\in\mathcal{C}^*} n_{\mathrm{C},ik} \quad \forall\, i \in \mathcal{M}^{(s)} \tag{C2}$$

$n_{\mathrm{C},ik}$ is the number of carbon atoms for molecule i in carbon type k, which is summed over detectable carbon types $\mathcal{C}^*$. n_{ig} is the number of FGs g in molecule i for the measured set $\mathcal{G}^*$. The carbon associated with carboxylic COOH is subtracted from $n^*_{\mathrm{C},i}$ before regression since $\lambda_{\mathrm{C,COOH}} \equiv 1$, and only aliphatic CH and alcohol aCOH is included in the fitting procedure. The detectable carbon fraction is estimated from the same mixtures by normalizing the abundance of detectable carbon over the total carbon (denoted by set $\mathcal{C}$):

$$\alpha^{(s)} = \left(\sum_{i\in\mathcal{M}^{(s)}}\sum_{k\in\mathcal{C}^*} n_{\mathrm{C},ik}\right) \Bigg/ \left(\sum_{i\in\mathcal{M}^{(s)}}\sum_{k\in\mathcal{C}} n_{\mathrm{C},ik}\right) \,.$$

$p(\lambda_{C,g})$ and $p(\alpha)$ are derived from the distribution of values estimated over realizations of subsets s.

Appendix D: Sampling the posterior distribution

Eq. 4 is typically posed as a mathematical problem to obtain the posterior distribution, written in this Section as $\pi(\theta) = p(\theta|y)$ for simplicity, from its unnormalized estimate $\tilde{\pi}(\theta) = p(y|\theta)p(\theta)$:

$$\pi(\theta) = \frac{1}{Z}\tilde{\pi}(\theta) = \frac{1}{Z}e^{-L(\theta)} \,. \tag{D1}$$

$L(\theta) = -\log\tilde{\pi}(\theta)$ is referred to as the loss function and Z is the normalizing constant (integral of $\tilde{\pi}(\theta)$ or $e^{-L(\theta)}$). In our model (eq. 1), we have both discrete and continuous parameters which we discriminate with superscripts (d) and (c), respectively. To explicitly expound on this notation, $\theta^{(c)} = \{\alpha, \kappa^2, \lambda_{C,g} : g \in \mathcal{G}^*\}$, $\theta^{(d)} = \{k_g : g \in \mathcal{G}^*\}$, and $\theta = \theta^{(c)} \cup \theta^{(d)}$. With $\theta'_i = \theta \setminus \{\theta_i\}$ denoting the set of all parameters except θ_i (i.e. the complement of θ_i with respect to θ), the marginal posterior distribution for θ_i is given by

$$\pi(\theta_i) = \frac{1}{Z}\sum_{\theta_i'^{(d)}} \int_{\theta_i'^{(c)}} \tilde{\pi}(\theta_i, \theta_i'^{(d)}, \theta_i'^{(c)}) d\theta_i'^{(c)} \,, \tag{D2}$$

with $\tilde{\pi}(\theta_i, \theta_i'^{(d)}, \theta_i'^{(c)}) = p(y|\theta_i, \theta_i'^{(d)}, \theta_i'^{(c)})p(\theta_i, \theta_i'^{(d)}, \theta_i'^{(c)})$. As with integral notation in eq. 4, the single integral or summation symbol applies over all parameters in the indexed set: i.e., $\int_\theta = \int_{\theta_1}\int_{\theta_2}\cdots\int_{\theta_D} \cdot d\theta_1 d\theta_2 \ldots d\theta_D$ and $\sum_{\theta_i'^{(d)}} = \sum_{\theta_{i,1}'^{(d)}}\sum_{\theta_{i,2}'^{(d)}}\cdots\sum_{\theta_{i,D(d)}'^{(d)}}$ $\int_\theta = \int_{\theta_1}\int_{\theta_2}\cdots\int_{\theta_{D^{(c)}}} \cdot d\theta_1 d\theta_2 \ldots d\theta_{D^{(c)}}$ and $\sum_{\theta_i'^{(d)}} = \sum_{\theta_{i,1}'^{(d)}}\sum_{\theta_{i,2}'^{(d)}}\cdots\sum_{\theta_{i,D^{(d)}}'^{(d)}}$. A summary of notation for posterior sampling is provided in Table A2. We use Markov Chain Monte Carlo (MCMC) as our primary tool to sample $\pi(\theta)$. To diagnose convergence and accuracy of the MCMC calculations, we additionally use a simple approximation (Laplace method) to confirm our parameter distributions. We first summarize Laplace method as it is a close extension of maximum likelihood estimation (MLE) typically used in conventional parameter estimation before describing MCMC sampling.

D1 Laplace method

The Laplace approximation (Tierney and Kadane, 1986; Murphy et al., 2012) solves eq. D1 and D2 by making a local Gaussian approximation to the posterior distribution of the continuous variables about their maximum a posteriori (MAP) estimate (i.e., maximum of the function $\tilde{\pi}$). This method improves on the classical MLE approach through the weighting of a prior (for a flat prior, the MAP estimate is equivalent to the MLE estimate), and estimating probabilities from the surface curvature of eq. D1 in the vicinity of the MAP. The approximation only applies in the domain of continuous parameters, so the calculation is performed for every selected realization of discrete parameter combinations. The probability estimate is formulated from the normalization constant of a multivariate normal distribution, with $D^{(c)} \times D^{(c)}$ Hessian $H_{\theta^{(c)*}}$ of L about $\theta^{(c)*}$:

$$\pi(\theta^{(c)}, \theta^{(d)}) = \left[\frac{\det H_{\theta^{(c)*}}}{(2\mathrm{pi})^{D^{(c)}}}\frac{\det H_{\theta^{(c)*}}}{(2\pi)^{D^{(c)}}}\right]^{1/2} e^{-\left[L(\theta^{(c)}, \theta^{(d)}) - L(\theta^{(c)*}, \theta^{(d)})\right]} \quad \forall\theta^{(d)} \,. \tag{D3}$$

Laplace's method is typically associated with a second-order Taylor series expansion about $\theta^{(c)*}$ which further provides the approximation: $L(\theta^{(c)},\theta^{(d)}) - L(\theta^{(c)*},\theta^{(d)}) \approx \frac{1}{2}(\theta^{(c)} - \theta^{(c)*})^T H_{\theta^{(c)*}}(\theta^{(c)} - \theta^{(c)*})$ for each realization of $\theta^{(d)}$. Covariance among the continuous variables can further be obtained from the inverse of the Hessian matrix. The marginal posterior for each realization of the variable θ_i is obtained by a Gaussian approximation for each integral in eq. D2 and calculating the $D^{(c)}-1 \times D^{(c)}-1$ Hessian $H_{\theta'^{(c)*}_i}$ about the MAP defined as $\theta'^{(c)*}_i = \arg\max_{\theta'^{(c)}_i} \tilde{\pi}(\theta_i, \theta'^{(c)}_i, \theta'^{(d)}_i)$:

$$\pi(\theta_i) = \sum_{\theta'^{(d)}_i} \left[\frac{\det H_{\theta^{(c)*}}}{(2\mathrm{pi})\det H_{\theta'^{(c)*}_i}} \frac{\det H_{\theta^{(c)*}}}{(2\pi)\det H_{\theta'^{(c)*}_i}} \right]^{1/2} e^{-\left[L(\theta_i,\theta'^{(c)*}_i,\theta'^{(d)}_i) - L(\theta^{(c)*},\theta^{(d)})\right]} . \tag{D4}$$

While analytically elegant and deterministic, the Laplace approximation is best suited for applications that primarily involve real (continuous) variables with a single mode in its probability density, or in the limit of large N as the density converges to a normal one (Bernstein-von Mises Theorem). However, its Gaussian estimates can become unreliable toward domain boundaries that might be imposed due to physical constraints, or in the limit of large number of variables when the high-dimensional space tends to become non-Gaussian.

We screen solutions by finding the MAP for each combination of discrete parameter values using L-BFGS-B (Limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm with box constraints), and removing those which are 10^{20} less than the absolute maximum. $\theta^{(c)*}$ for each realization of $\theta^{(d)}$ is found using L-BFGS-B, a box-constrained, limited-memory extension of the quasi-Newton method BFGS. BFGS uses an approximation of the Hessian matrix to steer its search. The Hessian matrix is not recomputed at each iteration but updated using the secant equation to account for the curvature estimated during the most recent step (Nocedal and Wright, 2006). While L-BFGS-B provides simultaneously provides estimation of the Hessian matrix with the MAP, as it is based on an approximation for the purposes of speeding up the optimization, we recompute these matrices and their determinants from numerical differentiation at the corresponding MAPs.

D2 MCMC

MCMC (Bishop, 2009; Aster et al., 2013) approximates the posterior probability $\pi(\theta)$ from an algorithmically-generated Markov sequence $\{\theta^{[1]}, \theta^{[2]}, \ldots, \theta^{[t]}, \ldots, \theta^{[n]}\}$. This sequence or chain is constructed through a series of trial and acceptance moves. The Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970) describes conditions under which the generated sequence fulfills the conditions of detailed balance necessary for convergence toward a stationary (statistically invariant) distribution. For any $\theta^{[t]}$, a candidate value θ^* is generated from a proposal distribution $q(\theta^*|\theta^{[t]})$. θ^* is designated as the next value in the sequence $\theta^{[t+1]}$ with acceptance probability $a(\theta^{[t]},\theta^*)$, defined to preserve detailed balance for a move from $\theta^{[t]}$ to θ^*:

$$a(\theta^{[t]},\theta^*) = \min\left\{1, \frac{q(\theta^{[t]}|\theta^*)\tilde{\pi}(\theta^*)}{q(\theta^*|\theta^{[t]})\tilde{\pi}(\theta^{[t]})}\right\} . \tag{D5}$$

The ratio $\tilde{\pi}(\theta^*)/\tilde{\pi}(\theta^{[t]})$ has been used in place of $\pi(\theta^*)/\pi(\theta^{[t]})$ so that explicit evaluation of the normalization constant Z (eq. D1) is not required. For a symmetric proposal distribution, $q(\theta^{[t]}|\theta^*) = q(\theta^*|\theta^{[t]})$ and further simplification to eq. D5 can be

obtained (Metropolis algorithm). Assignment of $\theta^{[t+1]}$ is implemented by comparison of $a(\theta^{[t]}, \theta^*)$ against the realization u of a random variable uniformly distributed over $[0, 1]$:

$$\theta^{[t+1]} = \begin{cases} \theta^* & \text{if } a(\theta^{[t]}, \theta^*) > u \text{ and} \\ \theta^{[t]} & \text{otherwise.} \end{cases}$$

The initial value $\theta^{[0]}$ of the Metropolis-Hasting algorithm is set at the maximum a posterior (MAP) estimated for the Laplace method. Proposal distributions for the discrete parameters k_g are truncated normal distributions which bounds the range of possible values. For continuous variables, the covariance matrix Σ of the target distribution is estimated using the first iterations of sampling, after which efficient proposal distributions are defined (Gelman et al., 2013):

$$q(\theta^{[t]}|\theta^*) \sim \mathrm{N}(\theta^*, c^2\Sigma) \quad \text{where} \quad c^2 \approx 2.4/\sqrt{D}\,.$$

Two MCMC chains were run for each model, and convergence was monitored using chain trace plots and Gelman-Rubin diagnostics (Gelman and Rubin, 1992). The posterior probability distribution $p(\theta)$, marginal distributions $p(\theta_i)$, population statistics of θ (including covariances), and posterior predictive distributions (Section 3.2) are then calculated from the numerically sampled sequence.

The distribution-free approach of this technique makes it applicable to discontinuous, non-differentiable functions, solutions at constraint boundaries, and to smaller datasets where the limiting distribution need not be normal. Sampling across models for model selection can also be handled by a special case of Metropolis-Hastings — transdimensional or reversible jump MCMC — in which the number of parameters for each model can vary (Green, 1995; Gallagher et al., 2009). While candidate PLS solutions generated with a different k_g (eq. B1) can also be interpreted as different models, for this study, k_g is treated as a discrete tuning parameter for the PLS model corresponding to a fixed calibration set. The typical downside of MCMC is the high computational cost, as large number of samples are needed for convergence and to ensure that the parameters sampled non-independently can provide adequate characterization of the target density. Where possible, use of MCMC together with simpler methods to confirm results is recommended (Brooks et al., 2011).

Author contributions. ST, AMD, and SLS conceived of the project. CB wrote the code, performed simulations, and analyzed results. MR prepared calibration models and guidance on their use. ST and CB wrote the manuscript; AMD and JLH provided regular input on the analysis and further editing of the manuscript. ST provided overall supervision of the project.

Code availability. Code for posterior sampling by MCMC and Laplace approximation is available at https://gitlab.com/aprl/fgoc-bayes.

Competing interests. The authors declare no competing interests.

Acknowledgements. The authors would like to thank Prof. Anthony Davison for helpful suggestions regarding Bayesian statistics and the Electric Power Research Institute contract 10003745 for funding.

References

Adamson, A. W.: A Textbook of Physical Chemistry, Academic Press, 2nd edn., 1979.

Aiken, A. C., Decarlo, P. F., Kroll, J. H., Worsnop, D. R., Huffman, J. A., Docherty, K. S., Ulbrich, I. M., Mohr, C., Kimmel, J. R., Sueper, D., Sun, Y., Zhang, Q., Trimborn, A., Northway, M., Ziemann, P. J., Canagaratna, M. R., Onasch, T. B., Alfarra, M. R., Prevot, A. S. H., Dommen, J., Duplissy, J., Metzger, A., Baltensperger, U., and Jimenez, J. L.: O/C and OM/OC ratios of primary, secondary, and ambient organic aerosols with high-resolution time-of-flight aerosol mass spectrometry, Environmental Science & Technology, 42, 4478–4485, https://doi.org/10.1021/es703009q, 2008.

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.

Anderson, J. A. and Seyfried, W. D.: Determination of Oxygenated and Olefin Compound Types by Infrared Spectroscopy, Analytical Chemistry, 20, 998–1006, https://doi.org/10.1021/ac60023a002, 1948.

Aster, R. C., Borchers, B., and Thurber, C. H.: Parameter estimation and inverse problems., Academic Press, Waltham, MA, https://doi.org/10.1016/C2009-0-61134-X, 2013.

Bahadur, R., Uplinger, T., Russell, L. M., Sive, B. C., Cliff, S. S., Millet, D. B., Goldstein, A., and Bates, T. S.: Phenol Groups in Northeastern US Submicrometer Aerosol Particles Produced from Seawater Sources, Environmental Science & Technology, 44, 2542–2548, https://doi.org/10.1021/es9032277, 2010.

Bayes, T.: An essay towards solving a problem in the doctrine of chances, Philosophical Transactions of the Royal Society of London, 53, 370–418, https://doi.org/10.1098/rstl.1763.0053, 1763.

Bishop, C. M.: Pattern recognition and machine learning, Springer, New York, NY, 2009.

Blanchard, C. L., Hidy, G. M., Shaw, S., Baumann, K., and Edgerton, E. S.: Effects of emission reductions on organic aerosol in the southeastern United States, Atmos. Chem. Phys., 16, 215–238, https://doi.org/10.5194/acp-16-215-2016, 2016.

Boris, A. J., Takahama, S., Weakley, A. T., Debus, B. M., Fredrickson, C. D., Esparza-Sanchez, M., Burki, C., Reggente, M., Shaw, S. L., Edgerton, E. S., and Dillner, A. M.: Quantifying organic matter and functional groups in particulate matter filter samples from the southeastern United States, part I: Methods, Atmospheric Measurement Techniques Discussions, 2019, 1–39, https://doi.org/10.5194/amt-2019-144, 2019.

Brooks, S., Gelman, A., Jones, G., and Meng, X.: Handbook of Markov Chain Monte Carlo, Chapman & Hall/CRC Handbooks of Modern Statistical Methods, CRC Press, 2011.

Brown, R. J. C., Beccaceci, S., Butterfield, D. M., Quincey, P. G., Harris, P. M., Maggos, T., Panteliadis, P., John, A., Jedynska, A., Kuhlbusch, T. A. J., Putaud, J.-P., and Karanasiou, A.: Standardisation of a European measurement method for organic carbon and elemental carbon in ambient air: results of the field trial campaign and the determination of a measurement uncertainty and working range, Environmental Science: Processes & Impacts, 19, 1249–1259, https://doi.org/10.1039/C7EM00261K, 2017.

Budisulistiorini, S. H., Li, X., Bairai, S. T., Renfro, J., Liu, Y., Liu, Y. J., McKinney, K. A., Martin, S. T., McNeill, V. F., Pye, H. O. T., Nenes, A., Neff, M. E., Stone, E. A., Mueller, S., Knote, C., Shaw, S. L., Zhang, Z., Gold, A., and Surratt, J. D.: Examining the effects of anthropogenic emissions on isoprene-derived secondary organic aerosol formation during the 2013 Southern Oxidant and Aerosol Study (SOAS) at the Look Rock, Tennessee ground site, Atmospheric Chemistry and Physics, 15, 8871–8888, https://doi.org/10.5194/acp-15-8871-2015, 2015.

Burnham, K. and Anderson, D.: Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, Springer New York, 2003.

Calvetti, D. and Somersalo, E.: Inverse problems: From regularization to Bayesian inference, Wiley Interdisciplinary Reviews: Computational Statistics, 10, e1427, https://doi.org/10.1002/wics.1427, 2018.

Carlton, A. G., de Gouw, J., Jimenez, J. L., Ambrose, J. L., Attwood, A. R., Brown, S., Baker, K. R., Brock, C., Cohen, R. C., Edgerton, S., Farkas, C. M., Farmer, D., Goldstein, A. H., Gratz, L., Guenther, A., Hunt, S., Jaeglé, L., Jaffe, D. A., Mak, J., McClure, C., Nenes, A., Nguyen, T. K., Pierce, J. R., de Sa, S., Selin, N. E., Shah, V., Shaw, S., Shepson, P. B., Song, S., Stutz, J., Surratt, J. D., Turpin, B. J., Warneke, C., Washenfelder, R. A., Wennberg, P. O., and Zhou, X.: Synthesis of the Southeast Atmosphere Studies: Investigating Fundamental Atmospheric Chemistry Questions, Bulletin of the American Meteorological Society, 99, 547–567, https://doi.org/10.1175/BAMS-D-16-0048.1, 2018.

Chan, T. W., Huang, L., Banwait, K., Zhang, W., Ernst, D., Wang, X., Watson, J. G., Chow, J. C., Green, M., Czimczik, C. I., Santos, G. M., Sharma, S., and Jones, K.: Inter-comparison of elemental and organic carbon mass measurements from three North American national long-term monitoring networks at a co-located site, Atmospheric Measurement Techniques, 12, 4543–4560, https://doi.org/https://doi.org/10.5194/amt-12-4543-2019, 2019.

Cheng, Y., Duan, F.-k., He, K.-b., Zheng, M., Du, Z.-y., Ma, Y.-l., and Tan, J.-h.: Intercomparison of Thermal–Optical Methods for the Determination of Organic and Elemental Carbon: Influences of Aerosol Composition and Implications, Environmental Science & Technology, 45, 10 117–10 123, https://doi.org/10.1021/es202649g, 2011.

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, 2011.

Chow, J. C., Watson, J. G., Chen, L.-W. A., Paredes-Miranda, G., Chang, M.-C. O., Trimble, D., Fung, K. K., Zhang, H., and Zhen Yu, J.: Refining temperature measures in thermal/optical carbon analysis, Atmos. Chem. Phys., 5, 2961–2972, https://doi.org/10.5194/acp-5-2961-2005, 2005.

Chow, J. C., Lowenthal, D. H., Chen, L.-W. A., Wang, X., and Watson, J. G.: Mass reconstruction methods for PM2.5: a review, Air Quality, Atmosphere & Health, 8, 243–263, https://doi.org/10.1007/s11869-015-0338-3, 2015.

Corrigan, A. L., Russell, L. M., Takahama, S., Äijälä, M., Ehn, M., Junninen, H., Rinne, J., Petäjä, T., Kulmala, M., Vogel, A. L., Hoffmann, T., Ebben, C. J., Geiger, F. M., Chhabra, P., Seinfeld, J. H., Worsnop, D. R., Song, W., Auld, J., and Williams, J.: Biogenic and biomass burning organic aerosol in a boreal forest at Hyytiälä, Finland, during HUMPPA-COPEC 2010, Atmospheric Chemistry and Physics, 13, 12 233–12 256, https://doi.org/10.5194/acp-13-12233-2013, 2013.

Davison, A. C. and Hinkley, D. V.: Bootstrap Methods and their Application, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge, https://doi.org/10.1017/CBO9780511802843, 1997.

Debus, B., Takahama, S., Weakley, A. T., Seibert, K., and Dillner, A. M.: Long-Term Strategy for Assessing Carbonaceous Particulate Matter Concentrations from Multiple Fourier Transform Infrared (FT-IR) Instruments: Influence of Spectral Dissimilarities on Multivariate Calibration Performance, Applied Spectroscopy, 73, 271–283, https://doi.org/10.1177/0003702818804574, 2019.

Decesari, S., Mircea, M., Cavalli, F., Fuzzi, S., Moretti, F., Tagliavini, E., and Facchini, M. C.: Source attribution of water-soluble organic aerosol by nuclear magnetic resonance spectroscopy, Environmental Science & Technology, 41, 2479–2484, https://doi.org/10.1021/es061711l, 2007.

Dillner, A. M. and Takahama, S.: Predicting ambient aerosol thermal-optical reflectance (TOR) measurements from infrared spectra: organic carbon, Atmospheric Measurement Techniques, 8, 1097–1109, https://doi.org/10.5194/amt-8-1097-2015, 2015.

Domingos, P.: A Few Useful Things to Know About Machine Learning, Communications of the ACM, 55, 78–87, https://doi.org/10.1145/2347736.2347755, 2012.

Dowd, P.: Quantifying the Impacts of Uncertainty, pp. 349–373, Springer International Publishing, Cham, https://doi.org/10.1007/978-3-319-78999-6_18, https://doi.org/10.1007/978-3-319-78999-6_18, 2018.

Duarte, R. M. and Duarte, A. C.: NMR Studies of Organic Aerosols, vol. 92 of *Annual Reports on NMR Spectroscopy*, pp. 83 – 135, Academic Press, https://doi.org/https://doi.org/10.1016/bs.arnmr.2017.04.003, http://www.sciencedirect.com/science/article/pii/S0066410317300145, 2017.

El-Zanan, H. S., Lowenthal, D. H., Zielinska, B., Chow, J. C., and Kumar, N.: Determination of the organic aerosol mass to organic carbon ratio in IMPROVE samples, Chemosphere, 60, 485 – 496, https://doi.org/10.1016/j.chemosphere.2005.01.005, 2005.

El-Zanan, H. S., Zielinska, B., Mazzoleni, L. R., and Hansen, D. A.: Analytical Determination of the Aerosol Organic Mass-to-Organic Carbon Ratio, Journal of the Air & Waste Management Association, 59, 58–69, https://doi.org/10.3155/1047-3289.59.1.58, 2009.

Epstein, S. A., Blair, S. L., and Nizkorodov, S. A.: Direct Photolysis of a-Pinene Ozonolysis Secondary Organic Aerosol: Effect on Particle Mass and Peroxide Content, Environmental Science & Technology, 48, 11 251–11 258, https://doi.org/10.1021/es502350u, 2014.

Frank, N. H.: Retained Nitrate, Hydrated Sulfates, and Carbonaceous Mass in Federal Reference Method Fine Particulate Matter for Six Eastern U.S. Cities, Journal of the Air & Waste Management Association, 56, 500–511, https://doi.org/10.1080/10473289.2006.10464517, 2006.

Frossard, A. A. and Russell, L. M.: Removal of Sea Salt Hydrate Water from Seawater-Derived Samples by Dehydration, Environmental Science & Technology, 46, 13 326–13 333, https://doi.org/10.1021/es3032083, 2012.

Fuller, W. A.: Measurement Error Models, John Wiley & Sons, New York, NY, 1987.

Gallagher, K., Charvin, K., Nielsen, S., Sambridge, M., and Stephenson, J.: Markov chain Monte Carlo (MCMC) sampling methods to determine optimal models, model resolution and model choice for Earth Science problems, Marine and Petroleum Geology, 26, 525–535, https://doi.org/10.1016/j.marpetgeo.2009.01.003, 2009.

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

Gelman, A. and Rubin, D. B.: Inference from iterative simulation using multiple sequences., Statistical Science, 7, 457–472, https://doi.org/10.1214/ss/1177011136, 1992.

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.

Gilardoni, S., Liu, S., Takahama, S., Russell, L. M., Allan, J. D., Steinbrecher, R., Jimenez, J. L., De Carlo, P. F., Dunlea, E. J., and Baumgardner, D.: Characterization of organic ambient aerosol during MIRAGE 2006 on three platforms, Atmospheric Chemistry and Physics, 9, 5417–5432, https://doi.org/10.5194/acp-9-5417-2009, 2009.

Green, P. J.: Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, Biometrika, 82, 711–732, https://doi.org/10.1093/biomet/82.4.711, 1995.

Griffiths, P. and Haseth, J. A. D.: Fourier Transform Infrared Spectrometry, John Wiley & Sons, In, 2nd edn., 2007.

Hand, J., Prenni, A., Schichtel, B., Malm, W., and Chow, J.: Trends in remote PM2.5 residual mass across the United States: Implications for aerosol mass reconstruction in the IMPROVE network, Atmospheric Environment, 203, 141 – 152, https://doi.org/10.1016/j.atmosenv.2019.01.049, 2019.

Hastie, T., Tibshirani, R., and Friedman, J.: The elements of statistical learning: data mining, inference, and prediction, Springer Verlag, 2009.

Hastings, W. K.: Monte Carlo sampling methods using Markov chains and their applications, Biometrika, 57, 97–109, https://doi.org/10.1093/biomet/57.1.97, 1970.

Hawkins, L. N., Russell, L. M., Covert, D. S., Quinn, P. K., and Bates, T. S.: Carboxylic acids, sulfates, and organosulfates in processed continental organic aerosol over the southeast Pacific Ocean during VOCALS-REx 2008, Journal of Geophysical Research-atmospheres, 115, https://doi.org/10.1029/2009JD013276, 2010.

Henderson, B. H., Pinder, R. W., Crooks, J., Cohen, R. C., Carlton, A. G., Pye, H. O. T., and Vizuete, W.: Combining Bayesian methods and aircraft observations to constrain the $HO^{\cdot}$ + NO_2 reaction rate, Atmospheric Chemistry and Physics, 12, 653–667, https://doi.org/10.5194/acp-12-653-2012, 2012.

Hettiyadura, A. P. S., Jayarathne, T., Baumann, K., Goldstein, A. H., de Gouw, J. A., Koss, A., Keutsch, F. N., Skog, K., and Stone, E. A.: Qualitative and quantitative analysis of atmospheric organosulfates in Centreville, Alabama, Atmospheric Chemistry and Physics, 17, 1343–1359, https://doi.org/10.5194/acp-17-1343-2017, 2017.

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.

Hudson, P. K., Gibson, E. R., Young, M. A., Kleiber, P. D., and Grassian, V. H.: Coupled infrared extinction and size distribution measurements for several clay components of mineral dust aerosol, Journal of Geophysical Research: Atmospheres, 113, D01 201, https://doi.org/10.1029/2007JD008791, 2008.

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, 1997.

Kabanikhin, S. I.: Definitions and examples of inverse and ill-posed problems, Journal of Inverse and Ill-posed Problems, 16, https://doi.org/10.1515/JIIP.2008.019, 2008.

Kamruzzaman, M., Takahama, S., and Dillner, A. M.: Quantification of amine functional groups and their influence on OM/OC in the IMPROVE network, Atmospheric Environment, 172, 124–132, https://doi.org/10.1016/j.atmosenv.2017.10.053, 2018.

Krapf, M., El Haddad, I., Bruns, E., Molteni, U., Daellenbach, K., Prévôt, A. H., Baltensperger, U., and Dommen, J.: Labile Peroxides in Secondary Organic Aerosol, Chem, 1, 603–616, https://doi.org/10.1016/j.chempr.2016.09.007, 2016.

Kuzmiakova, A., Dillner, A. M., and Takahama, S.: An automated baseline correction protocol for infrared spectra of atmospheric aerosols collected on polytetrafluoroethylene (Teflon) filters, Atmospheric Measurement Techniques, 9, 2615–2631, https://doi.org/10.5194/amt-9-2615-2016, 2016.

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.

Liu, S., Shilling, J. E., Song, C., Hiranuma, N., Zaveri, R. A., and Russell, L. M.: Hydrolysis of Organonitrate Functional Groups in Aerosol Particles, Aerosol Science and Technology, 46, 1359–1369, https://doi.org/10.1080/02786826.2012.716175, 2012.

Malm, W. C. and Hand, J. L.: An examination of the physical and optical properties of aerosols collected in the IMPROVE program, Atmospheric Environment, 41, 3407–3427, https://doi.org/10.1016/j.atmosenv.2006.12.012, 2007.

Malm, W. C., Sisler, J. F., Huffman, D., Eldred, R. A., and Cahill, T. A.: Spatial and seasonal trends in particle concentration and optical extinction in the United States, Journal of Geophysical Research: Atmospheres, 99, 1347–1370, https://doi.org/10.1029/93JD02916, 1994.

Marais, E. A., Jacob, D. J., Turner, J. R., and Mickley, L. J.: Evidence of 1991–2013 decrease of biogenic secondary organic aerosol in response to SO_2 emission controls, Environmental Research Letters, 12, 054 018, https://doi.org/10.1088/1748-9326/aa69c8, 2017.

Maria, S. F., Russell, L. M., Turpin, B. J., Porcja, R. J., Campos, T. L., Weber, R. J., and Huebert, B. J.: Source signatures of carbon monoxide and organic functional groups in Asian Pacific Regional Aerosol Characterization Experiment (ACE-Asia) submicron aerosol types, Journal of Geophysical Research-Atmospheres, 108, https://doi.org/10.1029/2003JD003703, 2003.

Martens, H. and Næs, T.: Multivariate Calibration, John Wiley & Sons, New York, 1991.

McClenny, W. A., Childers, J. W., Rōhl, R., and Palmer, R. A.: FTIR transmission spectrometry for the nondestructive determination of ammonium and sulfate in ambient aerosols collected on teflon filters, Atmospheric Environment, 19, 1891–1898, https://doi.org/10.1016/0004-6981(85)90014-9, 1985.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E.: Equation of State Calculations by Fast Computing Machines, The Journal of Chemical Physics, 21, 1087–1092, https://doi.org/10.1063/1.1699114, 1953.

Murphy, B. N., Donahue, N. M., Fountoukis, C., Dall'Osto, M., O'Dowd, C., Kiendler-Scharr, A., and Pandis, S. N.: Functionalization and fragmentation during ambient organic aerosol aging: application of the 2-D volatility basis set to field studies, Atmospheric Chemistry and Physics, 12, 10 797–10 816, https://doi.org/10.5194/acp-12-10797-2012, 2012.

Murphy, D. M., Cziczo, D. J., Froyd, K. D., Hudson, P. K., Matthew, B. M., Middlebrook, A. M., Peltier, R. E., Sullivan, A., Thomson, D. S., and Weber, R. J.: Single-particle mass spectrometry of tropospheric aerosol particles, Journal of Geophysical Research-atmospheres, 111, D23S32, https://doi.org/10.1029/2006JD007340, 2006.

Nocedal, J. and Wright, S. J.: Numerical Optimization, Springer, New York, NY, https://doi.org/10.1007/978-0-387-40065-5, 2006.

O'Hagan, T.: Dicing with the unknown, Significance, 1, 132–133, https://doi.org/10.1111/j.1740-9713.2004.00050.x, 2004.

Ott, W.: Environmental Statistics and Data Analysis, Taylor & Francis, 1994.

Pang, Y., Turpin, B. J., and Gundel, L. A.: On the Importance of Organic Oxygen for Understanding Organic Aerosol Particles, Aerosol Science and Technology, 40, 128–133, https://doi.org/10.1080/02786820500423790, 2006.

Pinder, R. W., Adams, P. J., Pandis, S. N., and Gilliland, A. B.: Temporally resolved ammonia emission inventories: Current estimates, evaluation tools, and measurement needs, Journal of Geophysical Research-atmospheres, 111, D16 310, https://doi.org/10.1029/2005JD006603, 2006.

Polidori, A., Turpin, B. J., Davidson, C. I., Rodenburg, L. A., and Maimone, F.: Organic PM2.5: Fractionation by polarity, FTIR spectroscopy, and OM/OC ratio for the Pittsburgh aerosol, Aerosol Science and Technology, 42, 233–246, https://doi.org/10.1080/02786820801958767, 2008.

Pope, R., Stanley, K. M., Domsky, I., Yip, F., Nohre, L., and Mirabelli, M. C.: The relationship of high PM2.5 days and subsequent asthma-related hospital encounters during the fireplace season in Phoenix, AZ, 2008–2012, Air Quality, Atmosphere & Health, 10, 161–169, https://doi.org/10.1007/s11869-016-0431-2, 2017.

Pye, H. O. T., Luecken, D. J., Xu, L., Boyd, C. M., Ng, N. L., Baker, K. R., Ayres, B. R., Bash, J. O., Baumann, K., Carter, W. P. L., Edgerton, E., Fry, J. L., Hutzell, W. T., Schwede, D. B., and Shepson, P. B.: Modeling the Current and Future Roles of Particulate Organic Nitrates in the Southeastern United States, Environmental Science & Technology, 49, 14 195–14 203, https://doi.org/10.1021/acs.est.5b03738, 2015.

Pye, H. O. T., D'Ambro, E. L., Lee, B. H., Schobesberger, S., Takeuchi, M., Zhao, Y., Lopez-Hilfiker, F., Liu, J., Shilling, J. E., Xing, J., Mathur, R., Middlebrook, A. M., Liao, J., Welti, A., Graus, M., Warneke, C., Gouw, J. A. d., Holloway, J. S., Ryerson, T. B., Pollack, I. B., and Thornton, J. A.: Anthropogenic enhancements to production of highly oxygenated molecules from autoxidation, Proceedings of the National Academy of Sciences, 116, 6641–6646, https://doi.org/10.1073/pnas.1810774116, 2019.

Ramadan, Z., Song, X.-H., and Hopke, P. K.: Identification of Sources of Phoenix Aerosol by Positive Matrix Factorization, Journal of the Air & Waste Management Association, 2011.

Ranney, A. P. and Ziemann, P. J.: Microscale spectrophotometric methods for quantification of functional groups in oxidized organic aerosol, Aerosol Science and Technology, 50, 881–892, https://doi.org/10.1080/02786826.2016.1201197, 2016.

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.

Reggente, M., Dillner, A. M., and Takahama, S.: Predicting ambient aerosol thermal-optical reflectance (TOR) measurements from infrared spectra: extending the predictions to different years and different sites, Atmospheric Measurement Techniques, 9, 441–454, https://doi.org/10.5194/amt-9-441-2016, 2016.

Reggente, M., Dillner, A. M., and Takahama, S.: Analysis of functional groups in atmospheric aerosols by infrared spectroscopy: systematic intercomparison of calibration methods for US measurement network samples, Atmospheric Measurement Techniques, 12, 2287–2312, https://doi.org/10.5194/amt-12-2287-2019, 2019.

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

Robert, C. P. and Casella, G.: Introducing Monte Carlo Methods with R, Springer Verlag, New York, https://doi.org/10.1007/978-1-4419-1576-4, 2010.

Robinson, A. L., Donahue, N. M., Shrivastava, M. K., Weitkamp, E. A., Sage, A. M., Grieshop, A. P., Lane, T. E., Pierce, J. R., and Pandis, S. N.: Rethinking organic aerosols: Semivolatile emissions and photochemical aging, Science, 315, 1259–1262, https://doi.org/10.1126/science.1133061, 2007.

Rock, D., Werts, C., Linn, R., and Joreskog, K.: A Maximum Likelihood Solution To The Errors In Variables And Errors In Equations Model, Multivariate Behavioral Research, 12, 187–197, https://doi.org/10.1207/s15327906mbr1202_6, 1977.

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, 2016.

Ruggeri, G., Bernhard, F. A., Henderson, B. H., and Takahama, S.: Model-measurement comparison of functional group abundance in α-pinene and 1,3,5-trimethylbenzene secondary organic aerosol formation, Atmospheric Chemistry and Physics, 16, 8729–8747, https://doi.org/10.5194/acp-16-8729-2016, 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.

Russell, L. M., Bahadur, R., and Ziemann, P. J.: Identifying organic aerosol sources by comparing functional group composition in chamber and atmospheric particles, Proceedings of the National Academy of Sciences of the United States of America, 108, 3516–3521, https://doi.org/10.1073/pnas.1006461108, 2011.

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.

San Martini, F. M., Dunlea, E. J., Volkamer, R., Onasch, T. B., Jayne, J. T., Canagaratna, M. R., Worsnop, D. R., Kolb, C. E., Shorter, J. H., Herndon, S. C., Zahniser, M. S., Salcedo, D., Dzepina, K., Jimenez, J. L., Ortega, J. M., Johnson, K. S., McRae, G. J., Molina, L. T., and Molina, M. J.: Implementation of a Markov Chain Monte Carlo method to inorganic aerosol modeling of observations from the MCMA-2003 campaign - Part II: Model application to the CENICA, Pedregal and Santa Ana sites, Atmospheric Chemistry and Physics, 6, 4889–4904, 2006.

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.

Schwarzenbach, R. P., Gschwend, P. M., and Imboden, D. M.: Environmental Organic Chemistry, John Wiley & Sons, 2nd edn., 2002.

Simon, H., Bhave, P. V., Swall, J. L., Frank, N. H., and Malm, W. C.: Determining the spatial and seasonal variability in OM/OC ratios across the US using multiple regression, Atmospheric Chemistry and Physics, 11, 2933–2949, https://doi.org/10.5194/acp-11-2933-2011, 2011.

Skoog, D., Holler, F., and Crouch, S.: Principles of Instrumental Analysis, Brooks/Cole Pub Co., Belmont, CA, 7th edn., 2017.

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, 2017.

Takahama, S., Liu, S., and Russell, L. M.: Coatings and clusters of carboxylic acids in carbon-containing atmospheric particles from spectromicroscopy and their implications for cloud-nucleating and optical properties, Journal of Geophysical Research-atmospheres, 115, D01 202, https://doi.org/10.1029/2009JD012622, 2010.

Takahama, S., Schwartz, R. E., Russell, L. M., Macdonald, A. M., Sharma, S., and Leaitch, W. R.: Organic functional groups in aerosol particles from burning and non-burning forest emissions at a high-elevation mountain site, Atmospheric Chemistry and Physics, 11, 6367–6386, https://doi.org/10.5194/acp-11-6367-2011, 2011.

Takahama, S., Johnson, A., Morales, J. G., Russell, L. M., Duran, R., Rodriguez, G., Zheng, J., Zhang, R., Toom-Sauntry, D., and Leaitch, W. R.: Submicron organic aerosol in Tijuana, Mexico, from local and Southern California sources during the CalMex campaign, Atmospheric Environment, 70, 500–512, https://doi.org/10.1016/j.atmosenv.2012.07.057, 2013a.

Takahama, S., Johnson, A., and Russell, L. M.: Quantification of Carboxylic and Carbonyl Functional Groups in Organic Aerosol Infrared Absorbance Spectra, Aerosol Science and Technology, 47, 310–325, https://doi.org/10.1080/02786826.2012.752065, 2013b.

Takahama, S., Dillner, A. M., Weakley, A. T., Reggente, M., Bürki, C., Lbadaoui-Darvas, M., Debus, B., Kuzmiakova, A., and Wexler, A. S.: Atmospheric particulate matter characterization by Fourier transform infrared spectroscopy: a review of statistical calibration strategies for carbonaceous aerosol quantification in US measurement networks, Atmospheric Measurement Techniques, 12, 525–567, https://doi.org/10.5194/amt-12-525-2019, 2019.

Thompson, R. L., Gerbig, C., and Rödenbeck, C.: A Bayesian inversion estimate of N_2O emissions for western and central Europe and the assessment of aggregation errors, Atmospheric Chemistry and Physics, 11, 3443–3458, https://doi.org/10.5194/acp-11-3443-2011, 2011.

Tierney, L. and Kadane, J. B.: Accurate Approximations for Posterior Moments and Marginal Densities, Journal of the American Statistical Association, 81, 82–86, https://doi.org/10.1080/01621459.1986.10478240, 1986.

Tukiainen, S., Railo, J., Laine, M., Hakkarainen, J., Kivi, R., Heikkinen, P., Chen, H., and Tamminen, J.: Retrieval of atmospheric CH4 profiles from Fourier transform infrared data using dimension reduction and MCMC, Journal of Geophysical Research: Atmospheres, 121, 10,312–10,327, https://doi.org/10.1002/2015JD024657, 2016.

Turpin, B. J. and Lim, H. J.: Species contributions to PM2.5 mass concentrations: Revisiting common assumptions for estimating organic mass, Aerosol Science and Technology, 35, 602–610, https://doi.org/10.1080/02786820152051454, 2001.

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, 2012.

Walter, E. and Pronzato, L.: Identification of Parametric Models from Experimental Data, Springer-Verlag, Berlin, 1997.

Wang, Y., Jiang, X., Yu, B., and Jiang, M.: A Hierarchical Bayesian Approach for Aerosol Retrieval Using MISR Data, Journal of the American Statistical Association, 108, 483–493, https://doi.org/10.1080/01621459.2013.796834, 2013.

Watson, J. G., Chow, J. C., and Chen, L.-W. A.: Summary of Organic and Elemental Carbon/Black Carbon Analysis Methods and Intercomparisons, Aerosol and Air Quality Research, 5, 65–102, https://doi.org/10.4209/aaqr.2005.06.0006, 2005.

Weisberg, S.: Applied Linear Regression, Wiley Series in Probability and Statistics, Wiley, 2013.

White, W. and Roberts, P.: On the nature and origins of visibility-reducing aerosols in the los angeles air basin, Atmospheric Environment (1967), 11, 803 – 812, https://doi.org/10.1016/0004-6981(77)90042-7, 1977.

Wold, S., Martens, H., and Wold, H.: The Multivariate Calibration-problem In Chemistry Solved By the PLS Method, Lecture Notes In Mathematics, 973, 286–293, 1983.

Zare, A., Fahey, K. M., Sarwar, G., Cohen, R. C., and Pye, H. O. T.: Vapor-Pressure Pathways Initiate but Hydrolysis Products Dominate the Aerosol Estimated from Organic Nitrates, ACS Earth Space Chem., 3, 1426–1437, https://doi.org/10.1021/acsearthspacechem.9b00067, 2019.

Figures

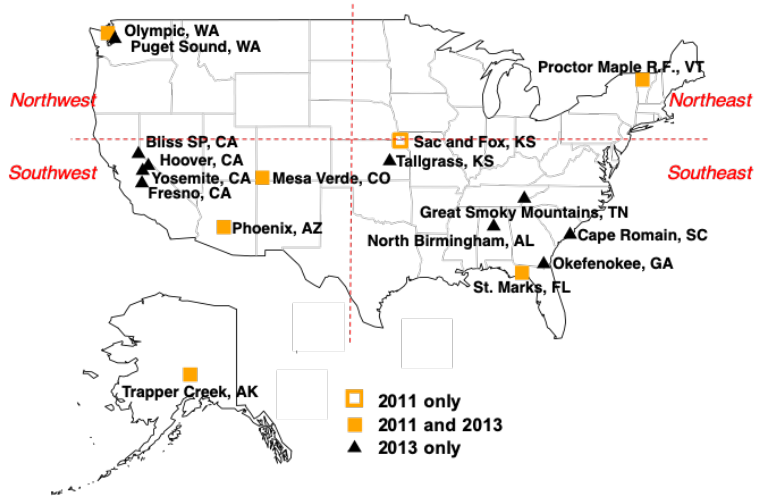

Figure 1. Map of IMPROVE network monitoring sites used in this work. For analysis in Section 4.4, the contiguous US is divided into four quadrants (vertical and horizontal red dashed lines centered at 40 °N and -100 °W); Alaska is considered as a separate region.

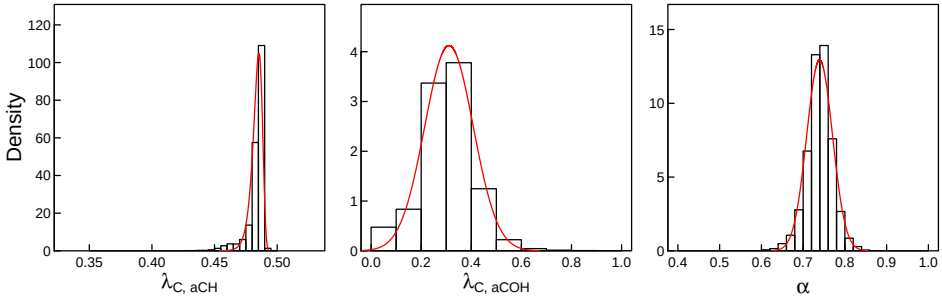

Figure 2. Prior distributions for λ_C and α. Histograms are generated from estimates from subsets of molecules ~~representating~~ representing a combination of primary and secondary organic aerosols, and red lines are fitted parametric distributions (Weibull for λ_C to capture asymmetry and normal for α).

~~Example posterior distribution of cluster 2 from MCMC. Dark lines correspond to prior distributions, blue histograms correspond to sampled posterior distributions, and red lines correspond to Laplace estimation~~

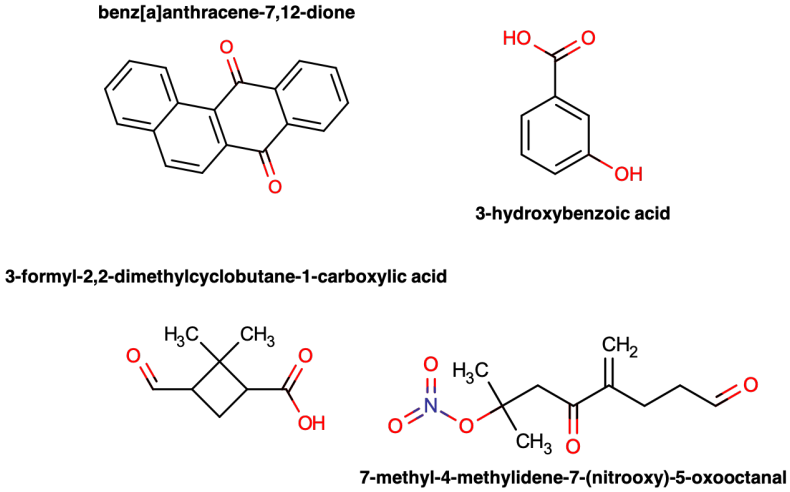

Figure 3. ~~Molecules~~ Examples of molecules containing carbon that are not detected by the measured set of FGs. ~~"C721CHO" and "C1010NO3" are names designated in the MCM v3.3.1 mechanism.~~

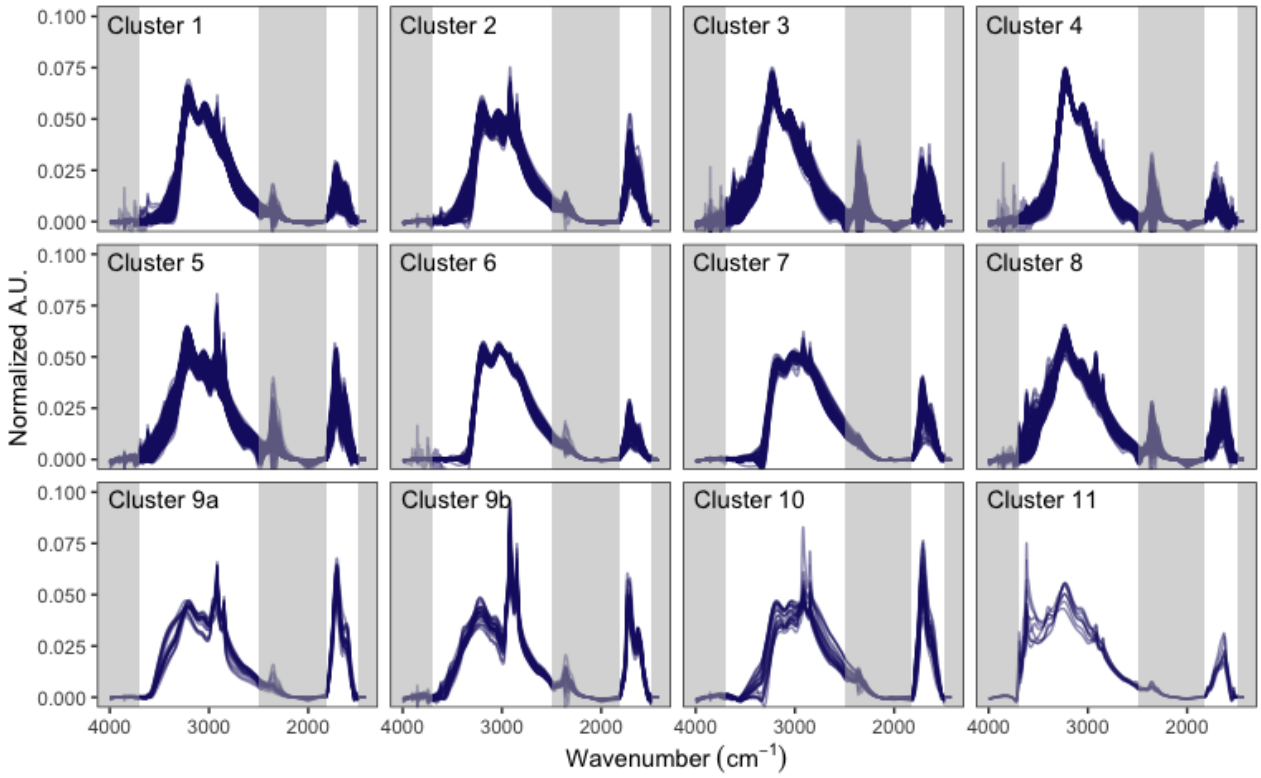

Figure 4. Visualization of spectral clusters. Gray vertical bars indicate regions excluded from cluster analysis. The clustering procedure and interpretation are described in Sections 3.1 and 4.2, respectively.

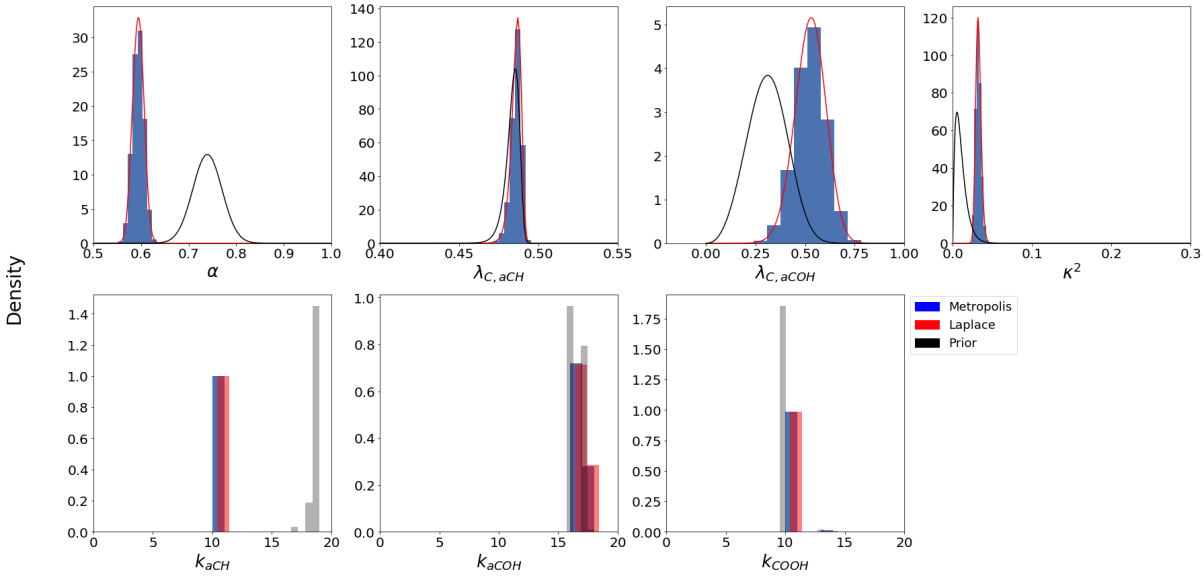

Figure 5. Example posterior distribution of cluster 2 from MCMC. Dark lines correspond to prior distributions, blue histograms correspond to sampled posterior distributions, and red lines correspond to Laplace estimation. "Density" refers to the probability or mass density and the variables are described in Sections 1.1 and 3.2. Non-parametric densities are approximated by kernel density estimation (Hastie et al., 2009) in figures.

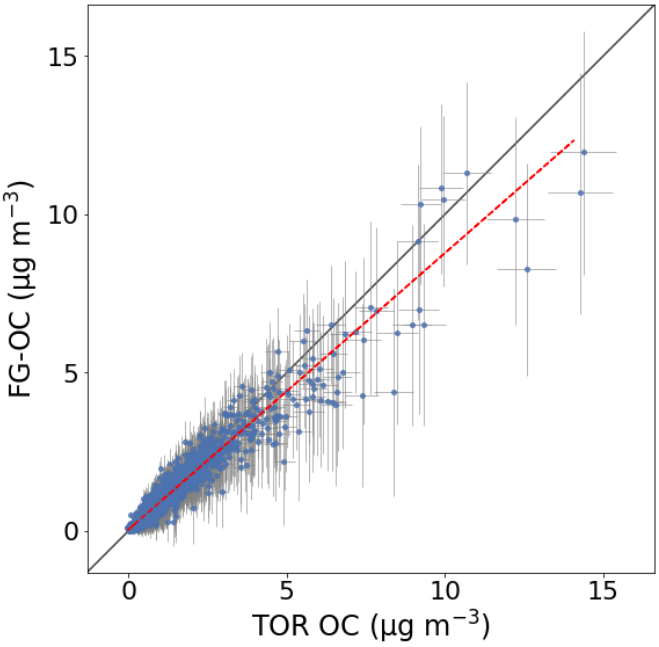

Figure 6. Comparison of reference TOR OC measurements and FG-OC estimated by Bayesian calibration. FG-OC corresponds to the mode of the posterior predictive distribution $\tilde{y}$ (Section S4). The lines span the 95% uncertainty intervals in TOR measurements horizontally, and 95% prediction intervals of the posterior distribution vertically. Diagonal line corresponds to 1:1 relation ~~for reference~~and the dotted red line coresonds to the best fit line (Pearson's r = 0.96, slope = 0.87, intercept = 0.04 μg m^{-3}).

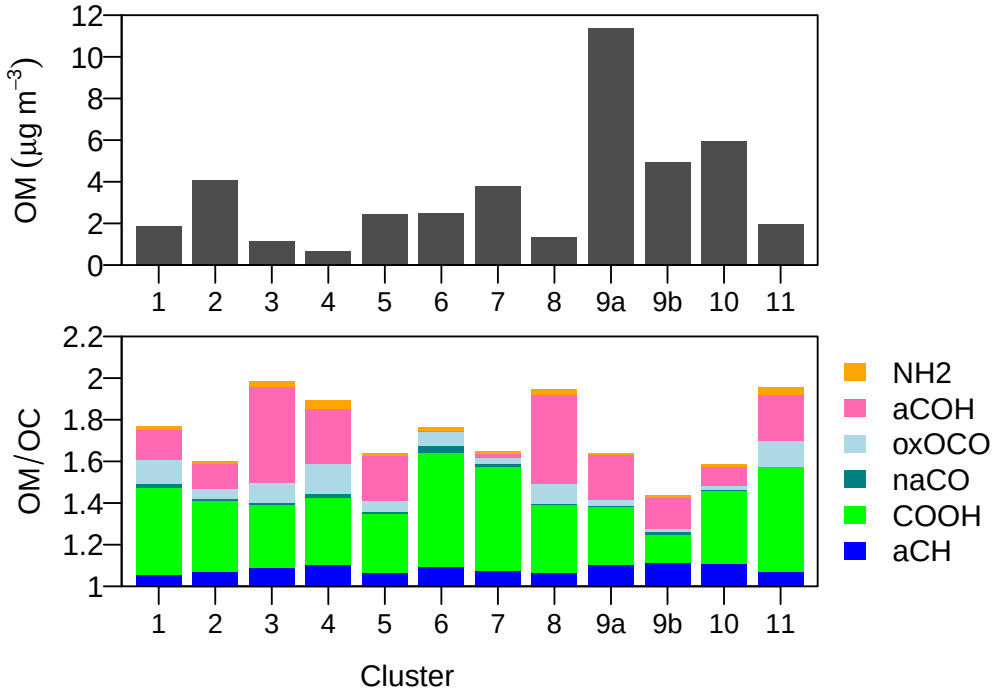

Figure 7. Mean OM and OM/OC for each cluster. Colors indicate FG contributions to the OM/OC.

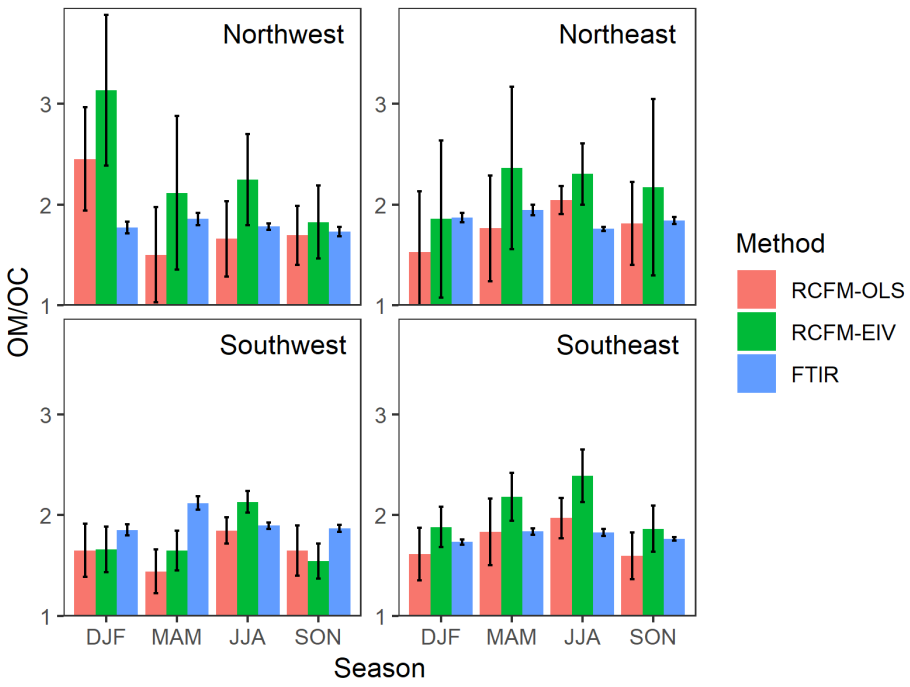

Figure 8. Estimates of OM/OC with 95 % confidence interval made by different techniques for the same sites for which FTIR measurements are available (Section 2). OLS (ordinary least squares) and EIV (error-in-variables) provide solutions to RCFM regression, and FTIR estimates are constructed from contributing functional groups. X-axes denote seasons: DJF (December, January, February) = winter, MAM (March, April, May) = spring, JJA (June, July, August) = summer, and SON (September, October, November) = fall.

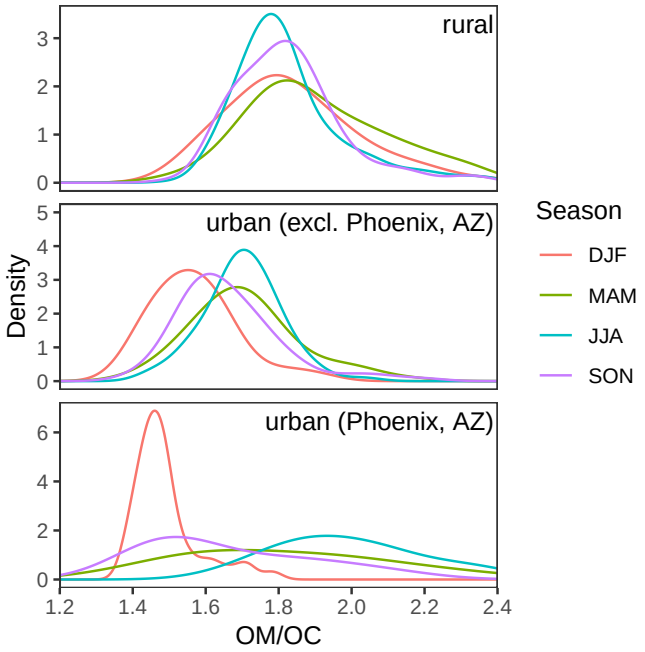

Figure 9. Probability densities of OM/OC estimated by FTIR for sites included in Figure 8, separated by site type. Densities for urban sites are separated into Phoenix, AZ, which is shown in its own panel, and the remaining five sites.

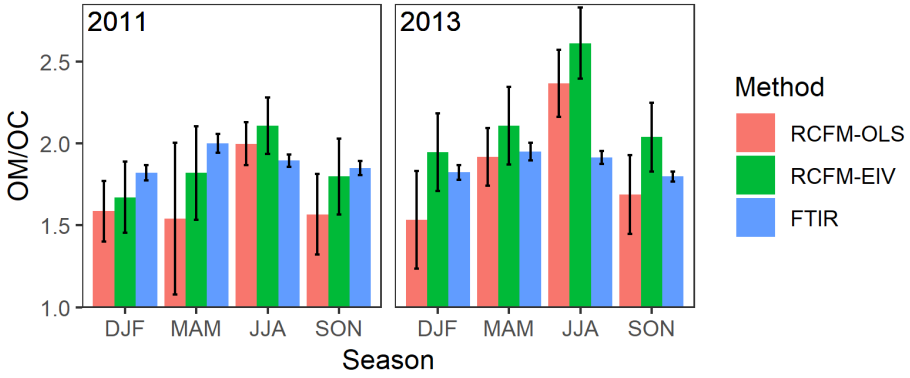

Figure 10. Estimates of OM/OC with 95 % confidence intervals for the same six sites for which FTIR measurements are available (one urban and six rural sites). The same notation as Figure 8 is used.

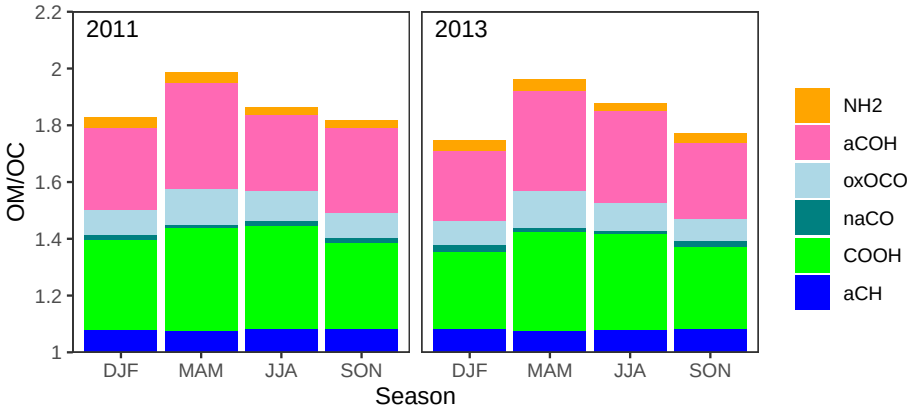

Figure 11. Mean OM/OC ratios partitioned by FG contributions for the FTIR estimates shown in Figure 10.

Tables

Table 1. Mode of parameter posterior distributions for each cluster.

Cluster	# samples	α	λ_{aCH}	λ_{aCOH}	k_{aCH}	k_{aCOH}	k_{COOH}	κ
1	387	0.59	0.49	0.44	7	16	10	0.20
2	176	0.60	0.49	0.53	10	16	10	0.18
3	771	0.81	0.43	0.59	13	12	10	0.27
4	442	0.83	0.48	0.07	16	17	10	0.31
5	343	0.57	0.49	0.44	10	16	10	0.17
6	87	0.80	0.48	0.58	10	16	10	0.20
7	68	0.66	0.49	0.59	10	16	10	0.20
8	128	0.71	0.48	0.50	10	16	10	0.29
9	43	0.76	0.48	0.37	16	16	10	0.13
10	21	0.79	0.48	0.21	16	16	10	0.13
11	8	0.71	0.49	0.32	9	16	10	0.17

Table A1. Notation for carbon estimation model.

Symbol	Description
n	moles (in areal density) of atom or functional group
x	infrared absorbance
λ	number of atoms per functional groups
α	carbon mass recovery fraction
m	mass of atom
M	atomic mass
t	PLS scores
p	PLS X-loadings
q	PLS Y-loadings
e	model residuals
k	number of latent variables in PLS model
$\mathcal{G}^*$	set of functional groups that are measured
$\mathcal{A}^*$	set of non-carbon atom types that are measured by $\mathcal{G}^*$
$\mathcal{C}$	set of carbon types
$\mathcal{C}^*$	set of carbon types that are measured by $\mathcal{G}^*$
n^*	moles (in areal density) of a unit measured by $\mathcal{G}^*$
$\mathcal{M}$	set of molecules
$\lvert\mathcal{M}\rvert$	number of molecules in set
ζ	fraction of primary to total (primary and secondary)

Table A2. Notation for Bayes theorem, likelihood, and posterior sampling algorithms.

Symbol	Description	Definition
y	data (observations); also outcome variable	TOR OC
θ	set of all parameters	$\theta^{(c)} \cup \theta^{(d)}$
$\theta^{(c)}$	set of continuous parameters	$\{\alpha, \lambda_{\mathrm{C,aCH}}, \lambda_{\mathrm{C,aCOH}}, \kappa^2\}$
$\theta^{(d)}$	set of discrete parameters	$\{k_{\mathrm{aCH}}, k_{\mathrm{aCOH}}, k_{\mathrm{COOH}}\}$
θ'_i	set of continuous parameters that excludes θ_i	$\theta \setminus \{\theta_i\} = \{\theta_i'^{(c)}, \theta_i'^{(d)}\}$
$\theta_i'^{(c)}$	set of continuous parameters that excludes θ_i	$\theta^{(c)} \setminus \{\theta_i\}$
$\theta_i'^{(d)}$	set of discrete parameters that excludes θ_i	$\theta^{(d)} \setminus \{\theta_i\}$
D	number of dimensions (parameters)	
p	probability density or mass function	
$\pi, \tilde{\pi}$	normalized and unnormalized posterior	
L	loss function	$\log \tilde{\pi}$
Z	normalizing constant	
H	Hessian matrix	
q	proposal distribution	
a	acceptance probability	

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 Prior distributions

This section includes Figures S1–S4; and Table S1.

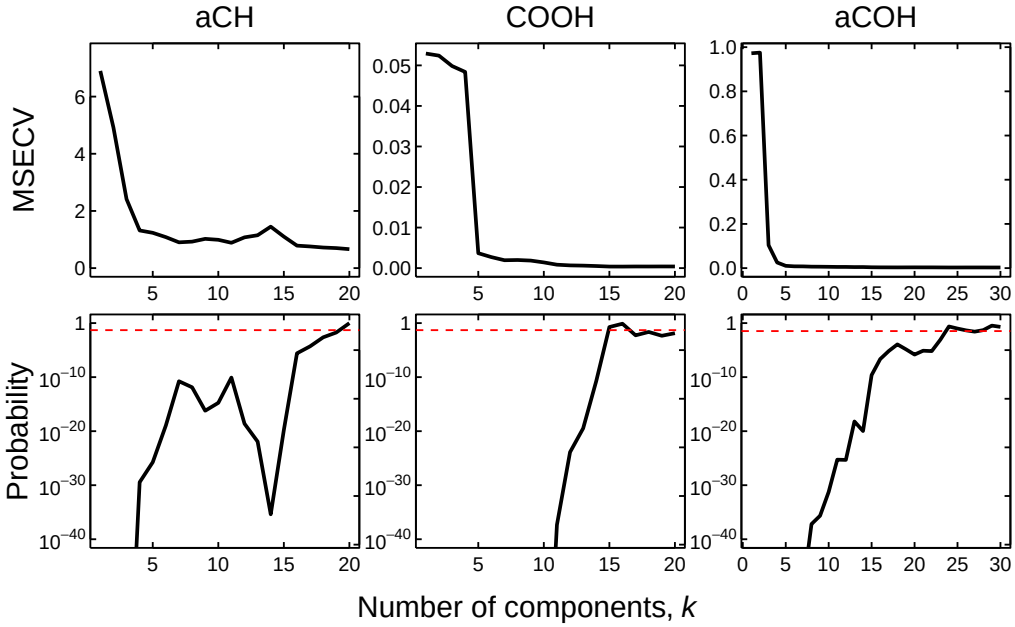

Figure S1. MSECV curves (in units of μmole of FG, top row) and resulting prior probability distributions for k (bottom row). Horizontal lines in in bottom row correspond to probability for a uniform distribution over the selected number of components.

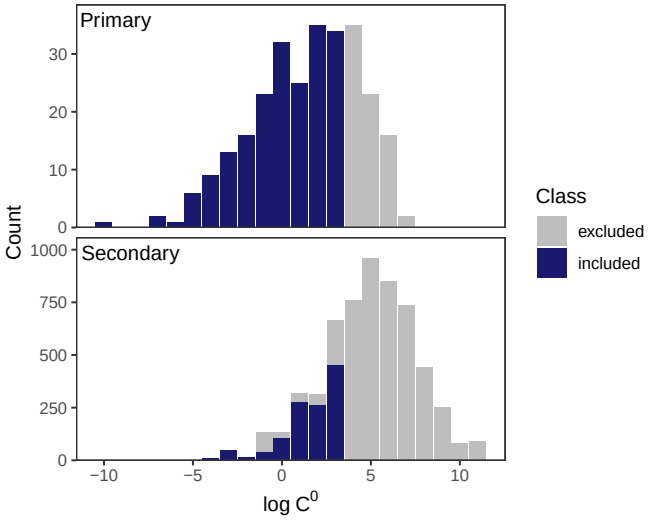

Figure S2. Distribution of equilibrium vapor concentrations C^0 ($\mu g\,m^{-3}$) for molecules taken from Rogge et al. (1993) and Rogge et al. (1998) ("Primary") and the MCM v3.3.1 database (Jenkin et al., 1997; Saunders et al., 2003) ("Secondary"). Only non-radical molecules with $C^0 \leq 10^{3.5}$ $\mu g\,m^{-3}$ are used in this study (excluded molecules below this threshold in the "Secondary" category represent radical species).

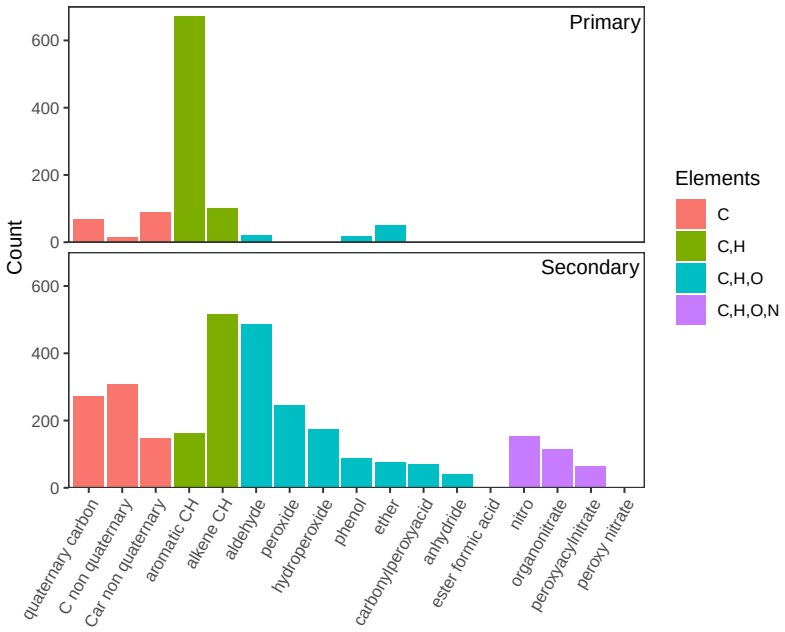

Figure S3. Number of molecular structures associated with undetected carbon atoms for all semivolatile compounds selected in Figure S2. Structures are colored by the elements that they contain. Structure names are described with illustrations in Table 1 of technical note by Ruggeri and Takahama (2016).

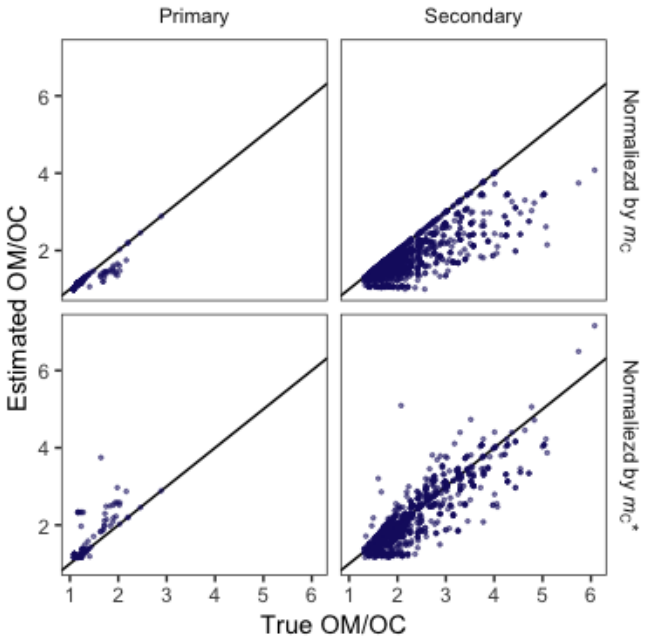

Figure S4. Estimates of OM/OC when normalized by m_C and αm_C. Secondary aerosol species contain many small but highly functional molecules, but the overall mode of the true OM/OC distribution is 1.96; the mode for primary aerosol species is 1.17.

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

Study	Mixture type	$\lambda_{C,CH}$	$\lambda_{C,aCOH}$
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.

S2 Contributions to the log-likelihood

In this section we outline calculations for assessing contribution of individual samples to the likelihood function (eq. 6). Let $r = (y - m_C)/\sigma$ 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 S5) 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 ($m_C \sim 3\sigma_0$) compared to one at the limit of quantification ($m_C \sim 10\sigma_0$) 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 $\kappa = 0.3$.

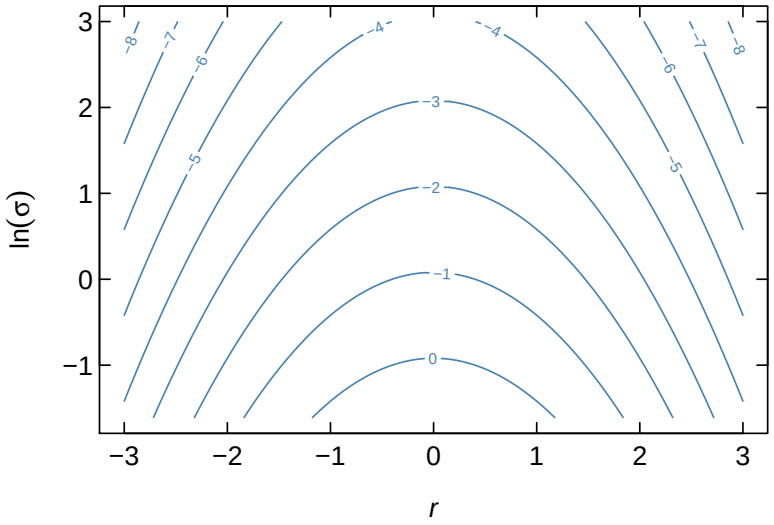

Figure S5. Isolines of $\ln f$ according to eq. S1.

S3 ~~Prior distributions~~Cluster analysis

~~This section includes Figures S1–S4; and Table S1.~~ Hierarchical cluster analysis (Bishop, 2009; Hastie et al., 2009) is used to categorize samples into spectroscopically similar groups (Russell et al., 2009; Liu et al., 2009; Ruthenburg et al., 2014). Spectra are first preprocessed by baseline correction (Kuzmiakova et al., 2016) and wavenumber selection (retaining only regions in the range 3700–2500 cm^{-1} and 1820–1500 cm^{-1}) to reduce the influence of substrate interference, particle scattering, and (carbon dioxide and water) vapors in the analysis chamber (Russell et al., 2009). The spectra are then normalized by their respective L2 norms (i.e., Euclidean distances from the origin when spectra are represented as vectors) so that they vary by composition rather than absolute absorbance (which includes the effect of mass loading in addition to composition). Finally, more than 1000 wavenumbers of the normalized spectra matrix are reduced to 9 dimensions using mean-centered, unscaled principal component analysis. These 9 principal components are selected from the eigenvalue profile ("scree plot") and their capability to explain 99% of the variance of the original spectra matrix. While instrumental noise does not contribute much to the overall signal (Debus et al., 2019), this preprocessing step additionally reduces the remaining water vapor contribution to the signal that is visible in spectra with low mass loadings, and makes distance metrics used for characterizing similarity more meaningful than what can be obtained in higher dimensions of correlated variables (Domingos, 2012).

~~MSECV curves (in units of 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.~~ The Euclidean distance metric with complete linkage is used for clustering samples based on their principal component scores. The number of clusters is heuristically selected by examining how the overall variability is reduced within each cluster (using the within sum-of-squares metric), and how well individual samples are served by the algorithmically-determined associations (with the Silhouette coefficient) with the creation of each additional cluster (Figure S6). Eleven superclusters are selected from this procedure, and model parameters θ estimated for each cluster and applied every member within it to predict FG-OC and FG-OM. As low signal-to-noise ratio samples can adversely affect the operations involving normalized spectra (i.e., principal component and cluster analyses), 10% of samples with the lowest L2 norms are initially excluded in the procedure above, but are assigned to the most appropriate cluster through k-nearest neighbor (k-NN) classification in the principal component space a posteriori for completeness.

~~Distribution of equilibrium vapor concentrations C^0 () 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 $C^0 \leq 10^{3.5}$ are used in this study (excluded molecules below this threshold in the "Secondary" category represent radical species).~~

~~Number of molecular structures associated with undetected carbon atoms for all semivolatile compounds selected in Figure S2. 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).~~

~~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.~~

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

S4 ~~Cluster analysis~~

This section includes Figures S6–S9.

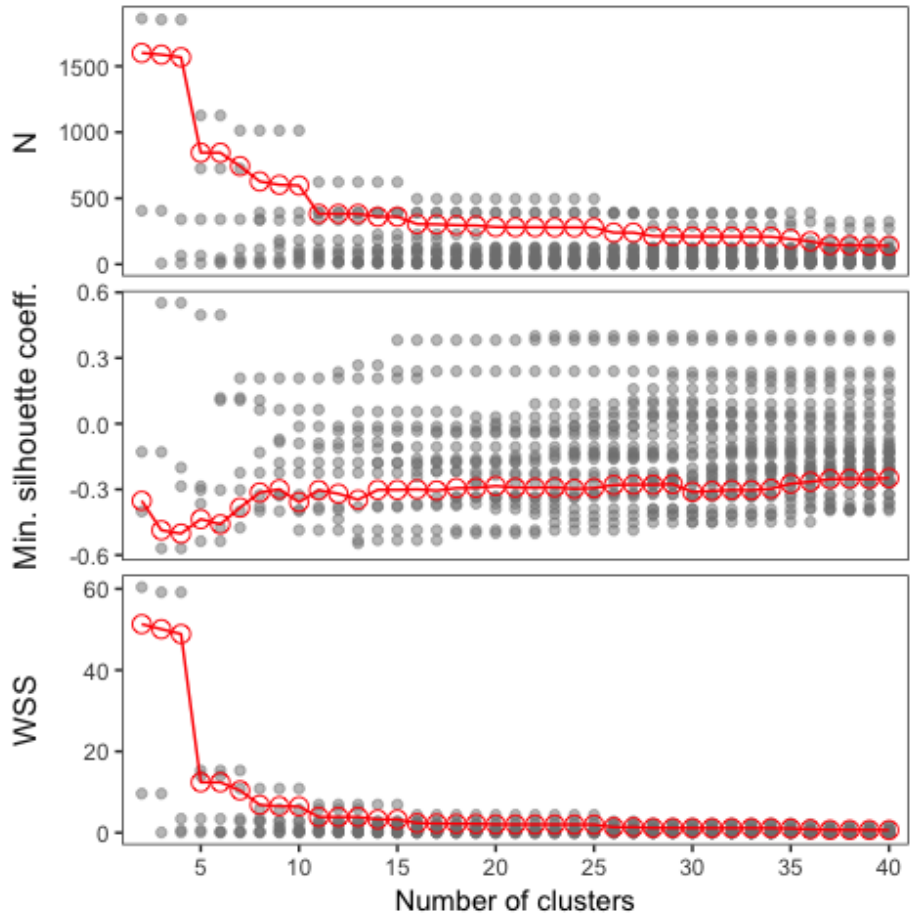

Figure S6. Number of samples in each cluster, minimum silhouette coefficient, and within sum-of-squares of each cluster as a function of the number of clusters formed. Gray points represent individual clusters, and red points and lines are values averaged across clusters.

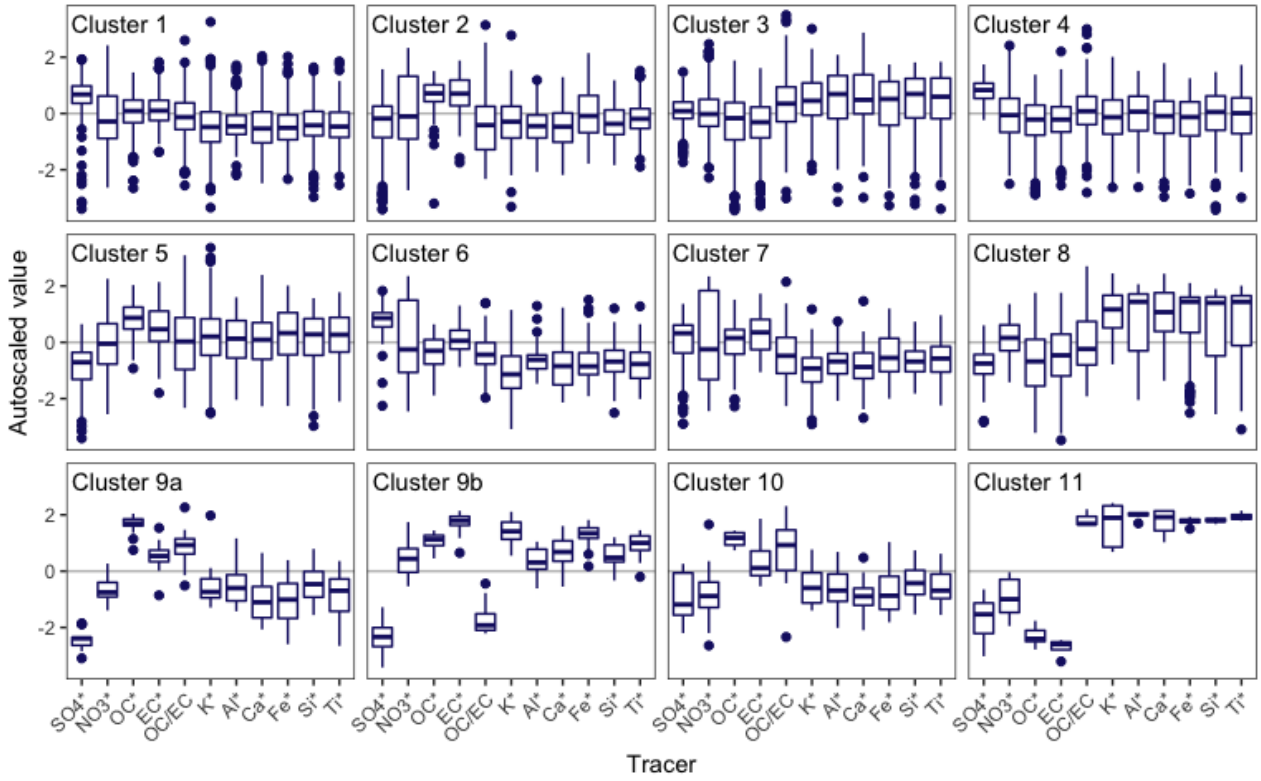

Figure S7. Comparisons of relative tracer concentrations. "*" denotes $PM_{2.5}$-normalized quantities. Normalized values are first logarithmically-transformed to be approximately symmetric, and then autoscaled (mean-centered and normalized by standard deviation of the variable for the entire data set). 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.

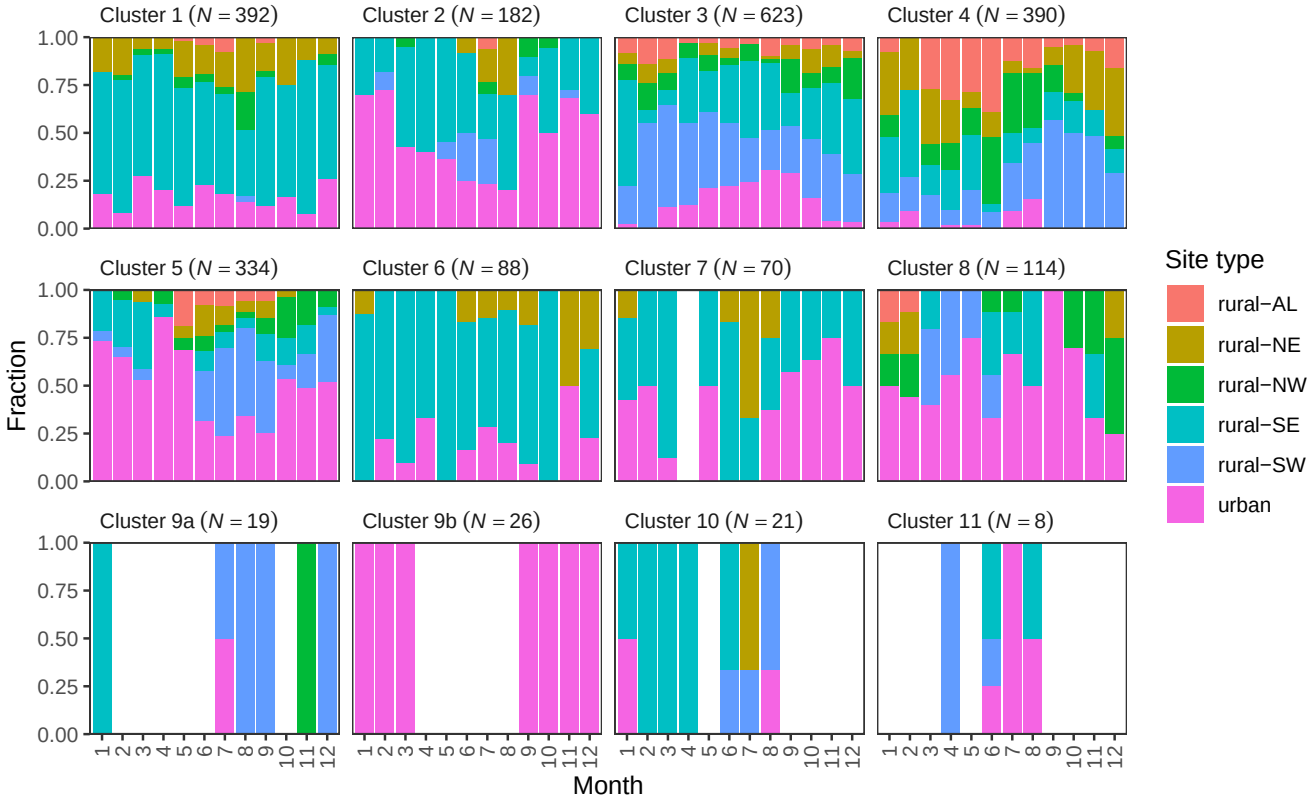

Figure S8. Composition of clusters.

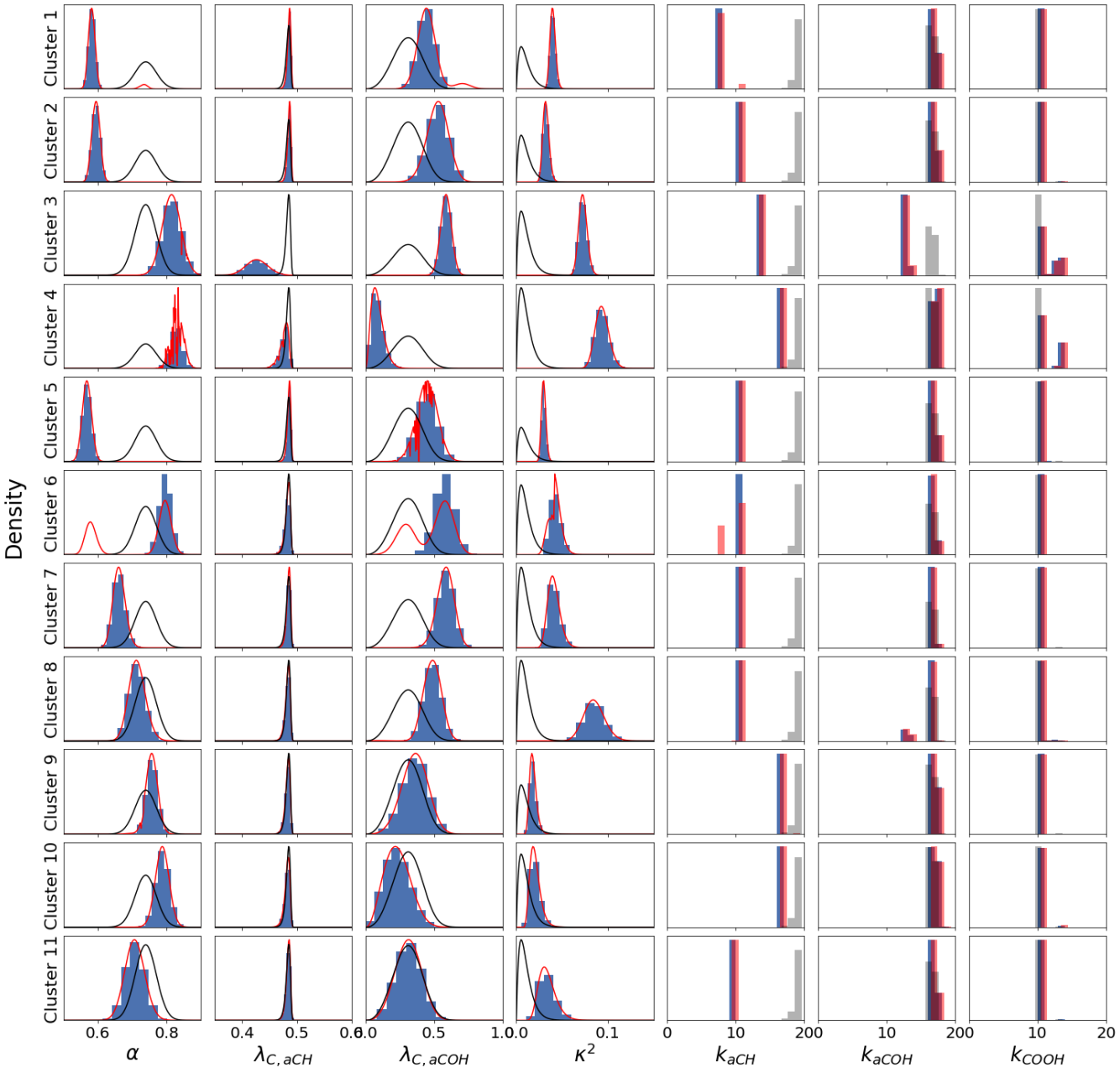

Figure S9. Posterior distributions of parameters for each cluster from MCMC (blue histograms) and L-BFGS-B (red lines) compared to prior distributions (black lines).

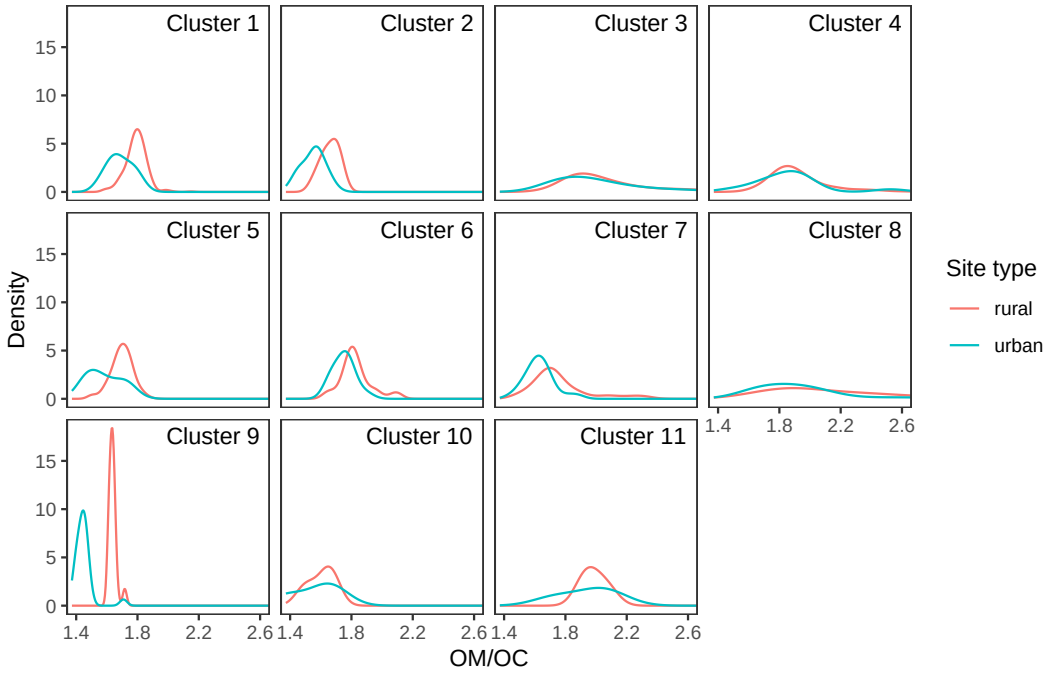

Figure S10. Probability distributions of OM/OC ratios segregated by site type.

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.

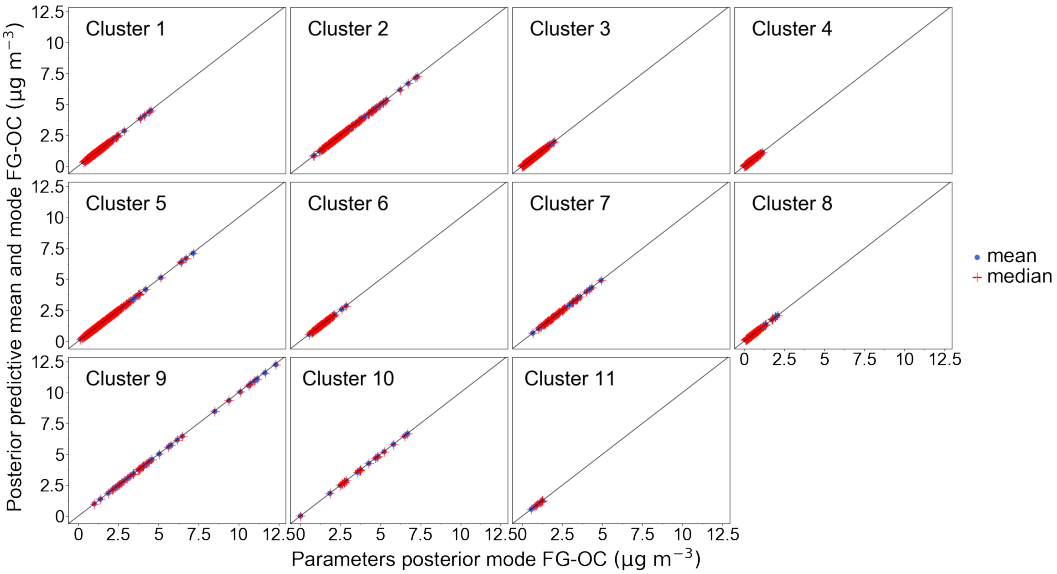

Figure S11. Comparison of central values of the posterior predictive distribution with predictions from single-point estimates of parameters obtained from their respective distributions.

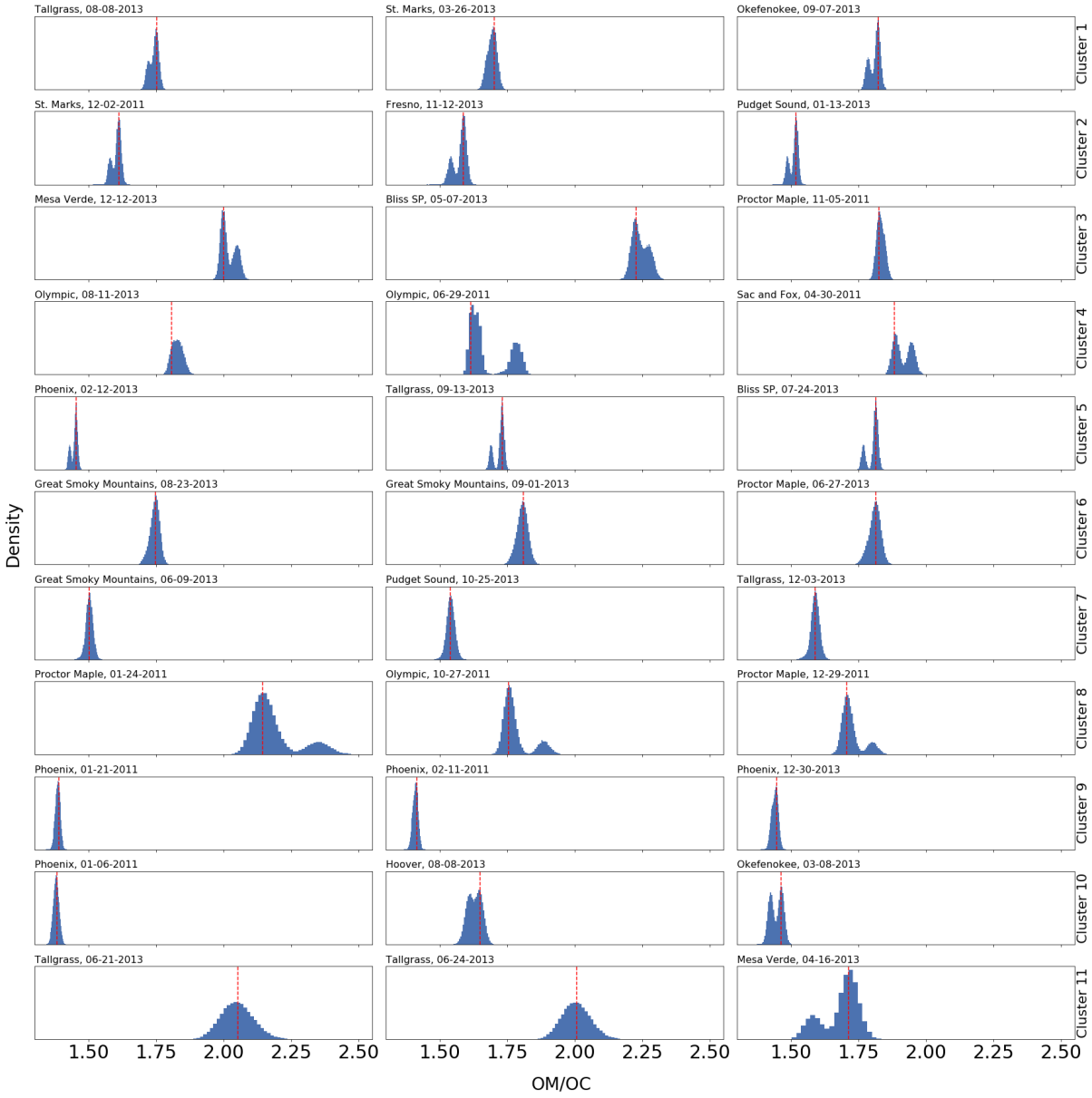

Figure S12. Posterior distributions for OM/OC. Red vertical lines indicate the reported value using MAP estimates of the parameters.

S5 Spatial and temporal prevalence of cluster types

This section includes Figures S13 and S14.

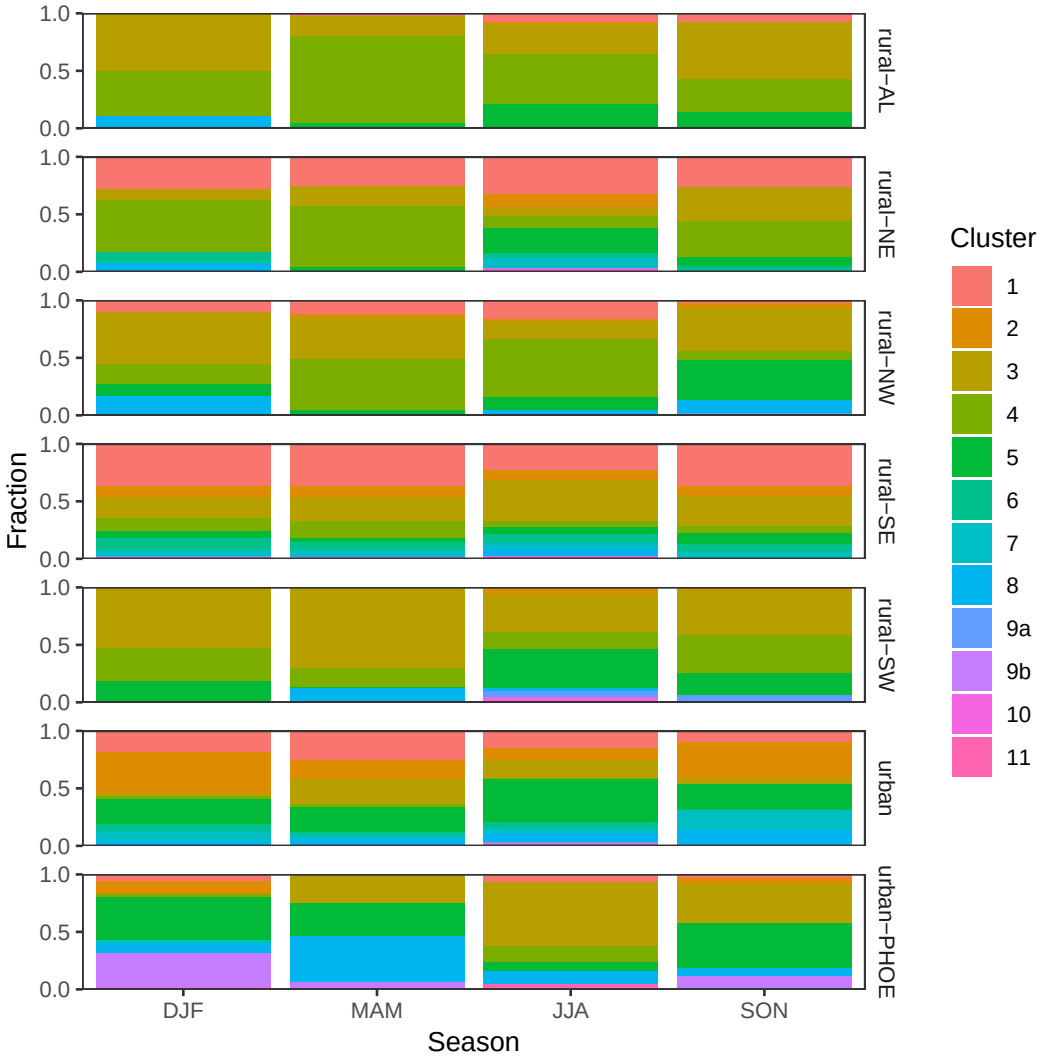

Figure S13. Frequency of clusters as fraction of samples at each site and season. "urban-PHOE" refers to Phoenix, AZ, and "urban" refers to all other urban sites.

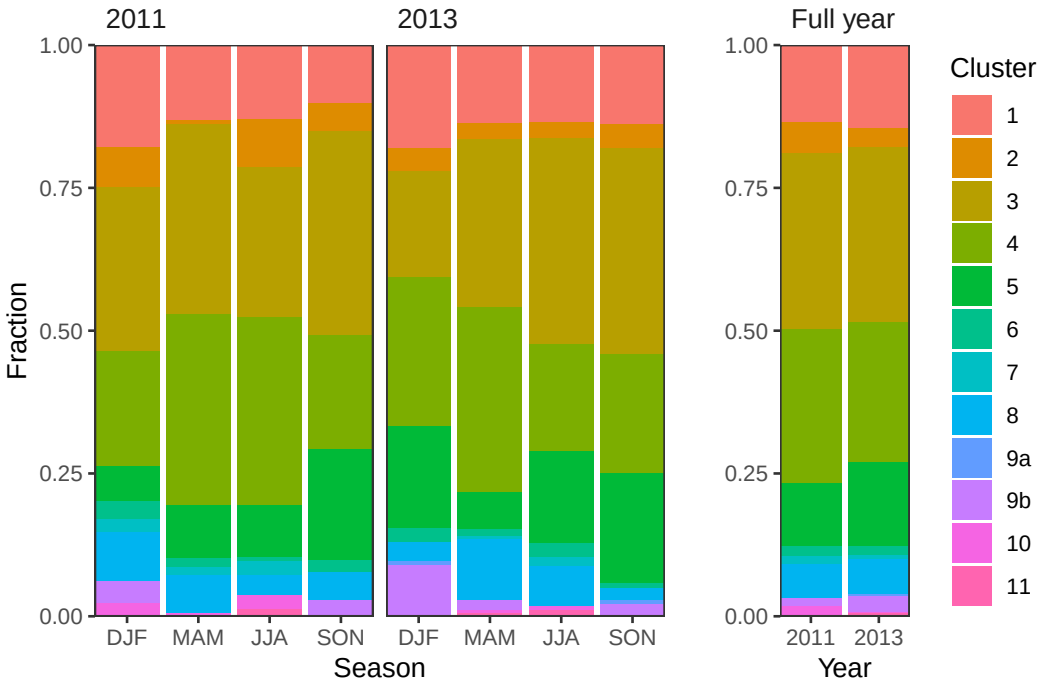

Figure S14. Frequency of clusters as fraction of samples during each year and season for six sites.

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.

Bishop, C. M.: Pattern recognition and machine learning, Springer, New York, NY, 2009.

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, 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.

Debus, B., Takahama, S., Weakley, A. T., Seibert, K., and Dillner, A. M.: Long-Term Strategy for Assessing Carbonaceous Particulate Matter Concentrations from Multiple Fourier Transform Infrared (FT-IR) Instruments: Influence of Spectral Dissimilarities on Multivariate Calibration Performance, Applied Spectroscopy, 73, 271–283, https://doi.org/10.1177/0003702818804574, 2019.

Domingos, P.: A Few Useful Things to Know About Machine Learning, Communications of the ACM, 55, 78–87, https://doi.org/10.1145/2347736.2347755, 2012.

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.

Hastie, T., Tibshirani, R., and Friedman, J.: The elements of statistical learning: data mining, inference, and prediction, Springer Verlag, 2009.

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, 1997.

Kuzmiakova, A., Dillner, A. M., and Takahama, S.: An automated baseline correction protocol for infrared spectra of atmospheric aerosols collected on polytetrafluoroethylene (Teflon) filters, Atmospheric Measurement Techniques, 9, 2615–2631, https://doi.org/10.5194/amt-9-2615-2016, 2016.

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. Non-catalyst 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, 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, 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, 2012.