Response to Reviewers

We thank the editor and anonymous reviewers for taking the time to oversee this process. Our point-by-point responses are included below in blue, and changes to the manuscript are highlighted in red.

Reviewer 1

General comments: "Atmospheric particulate matter characterization by Fourier Transform Infrared spectroscopy: a review of statistical calibration strategies for carbonaceous aerosol quantification in US measurement networks" addresses the need for accurate calibration models to interpret Fourier Transform Infrared (FT-IR) spectra, which can provide quantification of multiple species in complex atmospheric particulate matter (PM) mixtures. This review focuses on using models to facilitate quantitative predictions of organic carbon (OC) and elemental carbon (EC). The manuscript contributes substantial data thoroughly exploring several quantitative models for interpreting FT-IR spectra. The approaches and methods are valid and balanced, and maintain consistent self-checks for validity and bias.

We thank the reviewer for the assessment.

Specific comments:

1. The scope of the manuscript is very ambitious and covers detailed ground ranging from sample collection to model calibration maintenance. Potentially, the paper could be divided into two manuscripts. The manuscript starting from Background to the end of Section 3 could be its own review paper of methods and data-driven model-building, and Section 4 might be another review paper on operational phase models calibration and error prediction.

   We agree with the reviewer that this paper covers a lot of ground, but a large part of its value (currently lacking in the literature) is in bringing together different aspects of the calibration roadmap in a single source, and illustrating it with an extended example that is used across all parts.

2. The manuscript might benefit from a table that presents/overviews all the components explored in model building, evaluation, and interpretation.

   We have included a concise overview in Table C1.

3. The figures in general can be made larger to read the text more clearly. In particular, Figures 7 and 13 have very small text. Additionally, symbols can be used in Figure 10 to ensure readability in greyscale.

   We thank the reviewer for this comment. We have increased the font size in Figures 7 and 13 and changed the symbols in Figure 10 (and also Figure 3).

Technical corrections:

1. Page 9 line 23 manufacturer of the PTFE filters should probably say "Whatman". corrected
2. Page 19 line 3 abbreviation is "SAFOX" but Figure 6 uses the abbreviation "SAFO". should be SAFO
3. Page 23 line 9 should say, "A formal comparison. . . has not been performed, . . ." corrected
4. Page 24 line 12 should probably say "Wrappers operate under the implicit assumption. . ." corrected

5. Page 28 line 26 should probably say "While the large number of LVs used by the IMPROVE calibration models precluded attempts at identifying individual components. . ." corrected

We thank the reviewer for catching these errors.

Additional comments from Quick Review Report

1. Section 4.1 [operational phase] is not clearly connected back to Sections 3.4 and 3.5 (model interpretation and sample selection). For example, Section 3.4 points the reader to Section 4.1.2 [outlier detection] in particular, warning of the detriment in applying models in different contexts. However, Section 4.1.2 never again mentions the relationship between its content (outlier detection) and its impact on Section 3.4 (model interpretation). Additionally, while Section 4.1.2 Outlier detection makes a connection back to Section 3.5 Sample selection, it does not make the connection back to its impact on model interpretation. Perhaps when asking the reader in Section 3 to see information in Section 4, it can be restated in the beginning of Section 4 (or the appropriate subsection) why the reader might be interested in doing so.

   We thank the reviewer for this comment. We have made the connection more clear by adding the following text to Section 4 [operational phase]:

   "[T]his is the eventual use case for such calibration models — for instance, to enable FT-IR to provide TOR-equivalent carbon values from a PTFE filter in new monitoring sites or measurement campaigns where TOR analysis from a separate filter is not available. Without reference measurements, it is important to evaluate the appropriateness of available calibration models for new samples, continually monitor the performance of the model by introspective means, and update the calibration as necessary."

   Text in Section 4 [operational phase] now further ties back to Section 3.4 and 3.5 (new additions in italic):

   "Without reference measurements, many external indicators might be used to characterize differences between new samples and those in the calibration set*, especially with respect to attributes identified to be important (Section 3.5.1)*.

   [...]

   "Spectral matching combined with model interpretation (Section 3.4) can identify particular sample types that may be problematic for a calibration model *a priori*."

   and in Section 3.4 [interpretation] is changed to (new text in italic):

   "In particular, it is possible to exploit statistical correlations among the variables to make predictions, which can be detrimental if the correlation changes or model is applied in a different context ~~(further discussion is provided in Section 4.1.2)~~. Therefore, model interpretation is strongly related to anticipation of model applicability *and a priori identification of samples with potentially high prediction errors (Section 4.1.2)*."

2. The introduction of the manuscript is inadvertently misleading in how much it reviews laboratory-generated mixtures and simulated spectra. The introduction spends 2–3 (out of 4) pages reviewing methods and shortcomings for laboratory-generated mixtures and simulated spectra, which gives the impression that the bulk of the manuscript may also review laboratory standards and synthetic spectra. Perhaps the introduction can include subsection headers, such as "Limits of early applications" and "Data-driven approach."

   We thank the reviewer for the suggestion. We have added subsection headings, though we have named them "limits of conventional approaches to calibration" (Section 1.1) and "using collocated measurements" (Section 1.2) as some methods using laboratory standards can be considered to be data-driven, and are also currently in use. The use of collocated measurements provides a complementary approach, which we have now stated in the Introduction:

"[U]se of collocated measurements complement conventional approaches in expanding the capabilities of FT-IR spectroscopy to extract useful information contained in vibrational spectra."

Reviewer 2

General comments:

Takahama et al (2018), hereafter referred to as T2018 reviews methods used to determine quantitative measurements of OC/EC using FTIR spectroscopy on PTFE filter samples taken from sampling networks such as the Chemical Speciation Network (CSN) and the Interagency Monitoring of Protected Visual Environments (IMPROVE) network. The topic and scope of the article are very appropriate for AMT. There have been many papers on this topic in the past few years, so a review article summarizing the calibration, processing, and evaluation techniques is timely.

We thank the reviewer for the assessment.

1. However, it requires knowledge of the authors' previous papers, as well as other information. There are too many gaps in the information provided. Examples include such as

   - why PTFE is used and not quartz fiber filters

     We have added the following statements (in italics) in Section 2.1 [background on FT-IR]:

     "In this section, we cover the background necessary to understand FT-IR spectroscopy in the analysis of PM collected onto PTFE filter media*, which is optically thin and permits an absorbance spectrum to be obtained by transmission without additional sample preparation (McClenny et al., 1985; Maria et al., 2003).*"

     and in Section 2.2 [background on sample collection]:

     "*PTFE filters are used for gravimetric analysis on account of its low vapor absorption (especially water) and standardization in compliance monitoring, while quartz fiber filters are separately collected on account of its thermal stability (Chow, 1995; Chow et al., 2007b; Malm et al., 2011; Solomon et al., 2014; Chow et al., 2015).*"

   - the lack of general references for the measurement networks

     We have added references to Malm and Hand (2007) (IMPROVE) and Solomon et al. (2014) (CSN).

   - the lack of legends on the plots.

     We have included description of symbols and colors in the caption for space-constrained figures (which is a more traditional convention). We have modified the caption of Figure 7 to be more descriptive.

2. Additionally, there are some organizational problems, ranging from unclear section titles, to different datasets used in different ways, requiring effort from the reader to keep everything straight.

   We have renamed section titles to be more informative. For instance, "Interpretation" has been renamed to "Interpretation of important variables and their interrelationships" and "Model selection" to "Model selection without reference measurements" and "Calibration maintenance" to "Updating the calibration model". We have additionally clarified how the data is used; further clarification and corresponding changes to the paper are included in responses to specific comments 1–3.

3. Stated differently, this paper seems to focus on what the authors do and how they do it, which is important, but the authors need to emphasize why this is important, how it fits within the larger body of relevant literature, and make the paper stand on its own, while being clear to read.

We thank the reviewer for this perspective. In constructing this paper around a roadmap for statistical calibration, the importance of this paper is to provide generalization beyond what has been accomplished thus far for predicting TOR-equivalent carbon measurements and to enable extraction of useful information from vibrational spectra at an operational scale. This "in-situ" approach to calibration — using collocated measurements as reference — is an emerging strategy in atmospheric measurement research (e.g., for low-cost sensor calibration development) to increase the number of available measurements beyond that which more expensive reference measurements can provide. For the use case of FT-IR measurement of $PM_{2.5}$, the review pervasively draws upon the work of the authors in carbon quantification as an extensive example, as none other exist to our knowledge; a large number of citations are made with respect to the greater body of relevant literature in statistical calibration and statistical process control from which this framework is developed.

We have added restructured part of the introduction regarding calibration with collocated measurements and our demonstrated application to TOR analysis such that it reads:

"The benefit of building data-driven calibration models to reproduce concentrations reported by available measurements is twofold. One is to provide equivalent measurements when the reference measurements are expensive or difficult to obtain. For example, FT-IR spectra can be acquired rapidly, non-destructively, and at low cost from from Polytetrafluoroethylene (PTFE) filters commonly used for gravimetric mass analysis in compliance monitoring and health studies. That vibrational spectra contain many signatures of chemical constituents of PM (which also gives rise to challenges in spectroscopic interpretation) provides the basis for quantitative calibration of a multitude of substances. This capability for multi-analyte analysis is beneficial when a single filter may be relied upon during short-term campaigns, or in network sites for which installation of the full suite of instruments is prohibitive. The second benefit is the ability to gain a better understanding of atmospheric constituents measured by other techniques by associating them with important vibrational modes structural elements of molecules identified in the FT-IR calibration model. Such an application can be enlightening for studying aggregated metrics such as carbon content, or functional group composition in atmospheric PM quantified by techniques requiring more sample mass and user labor: ultraviolet-visible spectrometry or nuclear magnetic resonance spectroscopy (Decesari et al., 2003; Ranney and Ziemann, 2016).

In this paper, we demonstrate an extensive application of this approach in the statistical calibration of FT-IR spectra to collocated measurements of carbonaceous aerosol content — organic carbon (OC) and elemental carbon (EC) — characterized by a particular type of evolved gas analysis (EGA). EGA includes thermal optical reflectance (TOR) and thermal optical transmittance (TOT), which apportions total carbon into OC and EC fractions according to different criteria applied to the changing optical properties of the filter under stepwise heating (Chow et al., 2007a). EGA OC and EC are widely-measured in monitoring networks (Chow et al., 2007a; Brown et al., 2017), with historical significance in regulatory monitoring, source apportionment, and epidemiological studies. While EC is formally defined as sp^2-bonded carbon bonded only to other carbon atoms, what is measured by EGA EC is an operationally-defined quantity which is likely associated with low-volatility organic compounds (Chow et al., 2004; Petzold et al., 2013; Lack et al., 2014). EGA OC comprises a larger fraction of the total carbon and therefore less influenced by pyrolysis artifacts that affects quantification of EGA EC. In addition to OC estimates independently constructed from laboratory calibrations of functional groups, prediction of EGA OC and EC from FT-IR spectra will provide values for which strong precedent in atmospheric studies exist. Thus, use of collocated measurements complement conventional approaches in expanding the capabilities of FT-IR spectroscopy to extract useful information contained in vibrational spectra."

Specific Comments:

1. Site selection and use:

   - It's not always clear what datasets are used when, the map on Figure 1 suggests that the BYIS and FRES IMPROVE sites are used only as testing samples, but at the end of the paper the BYIS and FRES sites are used for training a new predictor.

     We have modified Figure 1 and its caption, and added the following statement to Section 2.2 [background on sample collection]:

     "TOR-equivalent carbon predictions for 2011 and 2013 IMPROVE samples discussed for this paper are made with a calibration model using a subset of samples from 2011 IMPROVE, and TOR predictions for 2013 CSN samples are made with a calibration model using a subset of samples from 2013 CSN. One exception is a special model constructed to illustrate how new samples can improve model prediction (Section 4.2); a subset of samples from two sites — Fresno, CA (FRES) and Baengnyeong Island, S. Korea (BYIS) — in 2013 IMPROVE are used to make predictions for the remaining samples at those sites. In all cases, analytical figures of merit for model evaluation are calculated for samples that are not used in calibration."

   - The 'data' section mentions calibration data from IMPROVE 2011, and CSN 2013, but it isn't clear that each of these datasets will be used separately and at different points within the paper. Thus, at various points in the paper: 2011 IMPROVE data are used to test calibrations using 2011 IMPROVE data, 2013 IMPROVE measurements are used to test a calibrations with 2011 IMPROVE data, 2013 IMPROVE data (For FRES and BYIS) are used to test calibrations with 2013 IMPROVE data, and 2013 CSN data are used to test calibrations with 2013 CSN data. In a paper this long, it can be challenging to remember what data is used where.

     The modification to the text in response to the point above also now explicitly states that subsets of 2011 IMPROVE and 2013 CSN data are used to build calibration models for IMPROVE and CSN predictions, respectively, with the exception of the special model using FRES and BYIS samples.

2. Some other site selection questions: A user might wonder why CSN and IMPROVE data are not used together to develop a model: there are some differences in the collection method between CSN and IMPROVE (Weakley et al., 2016), but no discussion of these differences is found in the paper. The Elisabeth (ELLA) required a special calibration. Weakley et al. (2016) noted that this site was located near a refinery but no discussion of the potential influence of the refinery on different spectra is discussed.

   We thank the reviewer for pointing out the potential to use IMPROVE and CSN data together for calibration. We have now addressed this point by adding the following statement in Section 2.2 [background on sample collection]:

   "Given the different sampling protocols that result in different spectroscopic interferences from PTFE (due to different filter types) and range of mass loadings (due to flowrates), and difference in expected chemical composition (due to site types), calibrations for the CSN and IMPROVE networks have been developed separately (Weakley et al., 2016). Advantages of building such specialized models in favor of larger, all-inclusive models are discussed in Section 3.5."

   Regarding Elizabeth, NJ, (ELLA) the site was located near a toll station in the NJ turnpike (not refinery) and the impact of potentially high levels of diesel PM are discussed in Section 3.4, to which we added the possible impact of the nearby source (in italics).

   "Weakley et al. (2018) found that a calibration model for ELLA did not require aromatic structures for prediction of TOR-equivalent EC. This site was *located in close proximity to a toll station on the New Jersey turnpike and was* characterized by high diesel PM loading, low OC/EC ratio, and low degree of

charring compared to samples from other CSN sites in the 2013 data set. The calibration model was able to predict TOR- equivalent EC concentrations primarily using absorption bands associated with aliphatic C-H (also selected in the calibration model for the other 2013 CSN sites) and nitrogenated groups believed to be markers for diesel PM."

3. Section titles could be improved. Section 4.2 is an example of this, a more descriptive title like "Applicability of calibrations developed under one set of conditions to samples measured under new conditions" would be more descriptive.

   Calibration maintenance is a technical phrase used in chemometrics but we have renamed the section to "Updating the calibration model" to be more descriptive; other changes to section titles are included in response to general comment 2.

4. One of the use cases for this technique is in a network where OC/EC measurements are not made using the standard EGA technique (Page 6, Line 30). The discussion in 4.2 [calibration maintenance] directly touches this concept - but the topic mentioned in the introduction isn't really brought up in that section.

   This is actually relevant for both Sections 4.1 [anticipating errors] and 4.2 [updating calibration]. We have changed the opening statement of Section 4 [operational phase] to include the italicized statement:

   "The operational phase of the model marks a departure from the building and evaluation phases (Figure 2) in that reference measurements may no longer be available on a regular basis. *However, this is the eventual use case for such calibration models — for instance, to enable FT-IR to provide TOR-equivalent carbon values from a PTFE filter in new monitoring sites or short-term field campaigns where collection and analysis of PM on a separate quartz fiber filter for TOR analysis is prohibitive. Without reference measurements, it is important to* continually monitor the performance of the model by introspective means, and update the calibration as necessary."

   And in Section 4.2 [updating calibration]:

   "In the context of FT-IR measurements, TOR reference measurements may not be available for short-term campaigns at new sites and some aspects of transfer learning and transductive learning strategies (sample reweighting or basis-set rederivation) may be the only option for improvement if prediction errors from existing calibration models are expected to be high (Section 4.1). For long-term operation at a fixed site, collecting a limited number of reference samples for recalibration initially or periodically can be a viable strategy if sample characteristics substantially differ from those available for calibration."

5. The data availability does not mention where to obtain the FTIR Spectra. Also, I could not find the source code at http://airspec.epfl.ch.

   The FT-IR data will be hosted in a publicly-accessible repository, but in the meantime can be obtained from the authors directly. The source code for AIRSpec is currently under review with another manuscript but can be obtained together with its underlying packages at https://aprl.epfl.ch/page-130782-en.html.

6. The authors should include a list of acronyms as an appendix.

   We have included Table B1 in the appendix that lists acronyms used in multiple sections.

7. There are no calibration sites in the Midwestern states (e.g. at longitudes between Birmingham, Alabama and Mesa Verde, CO), other than the Sac and Fox site which only has one half years' worth of data. Is this a problem for applicability of the model?

   The range of sites, local sources, and the meteorological conditions represented in the calibration samples are relevant only to the extent that they add to the diversity of chemical composition, which enables application of the model to new samples with similar composition. As we do not have many sites in the Midwest with which we can evaluate the model, presently it is difficult to determine whether

the current calibration models would suitable. If FT-IR spectra were available from Midwestern sites (without TOR measurements), error anticipation methods (Section 4.1) can be used together with laboratory calibrations (e.g., functional group measurements) to determine how similar the samples are spectroscopically and chemically to samples already present in the current calibration set (from mostly non-Midwestern states).

We have modified the text as included in response to comment 8 below.

8. The authors briefly mention meteorological influences on the calibrations – could model error be used to infer something about the variability due to meteorological conditions? Are there better performances across the different months – e.g. both meteorological patterns, as well as well as combustion patterns, are different between summer and winter. How does calibration data during only a short period (such as at the Sac and Fox site) bias the results?

   The variability due to meteorology and other environmental conditions are only relevant to the extent that they change the sample composition outside of the range encountered in the calibration set. In Section 4 [operational phase], we summarize how Reggente et al. (2016) found that calibration developed under one year was able to provide predictions for another year (and also at different sites), but this is possibly because variations across all seasons were represented in the calibration set. Using calibration from a short period can bias predictions if the range of composition encountered over subsequent periods (e.g., different seasons) are not well-represented during the short period of calibration. This is an area that can benefit from complementary spectral analysis to assess variability (e.g., in terms of functional group composition).

   We have modified Section 3.2.2 [model evaluation] with italics indicating new additions:

   "For instance, high prediction errors elevated over multiple days may be associated with aerosols of *unusual* composition transported under synoptic scale meteorology *that is not well-represented in the calibration samples*."

   In Section 3.5 [sample selection], regarding use of the stratified sample selection approach (selecting samples spaced out over one year at each measurement site), the following text has been added:

   "[S]amples from the same site and season are not strictly required for successful prediction of each new sample. Reggente et al. (2016) demonstrate accurate prediction for a full year of TOR OC and EC concentrations at sites not included in the calibration (also revisited in Section 4.1). The extent to which site, season, local emission, or meteorological regime of a new sample affects prediction depends on how these factors contribute to deviation in chemical composition from calibration samples."

   And in Section 4.1.2 [outlier detection], the italicized statement has been added:

   "[T]he actual increase in prediction error (if any) will depend on the functional relationship among variables and how well they are represented by the model — e.g., a linear relationship modeled by a linear mapping may perform adequately in interpolation and extrapolation. *For instance, samples with OM/OC and OC/EC composition and TOR OC concentrations out of range with respect to calibration samples were predicted without substantial increase in errors (Section 3.5.1).* Therefore, not all outliers may be associated with high prediction errors."

Technical corrections:

1. P3L25: smog chamber → smog chambers corrected
2. P6L23: measured by EGA EC is an → measured by EGA, EC is an corrected to "EC measured by EGA"
3. P6L25: and therefore less influenced → and therefore is less influenced corrected

4. P9L31: "Change to artifact correction method for OC carbon fractions" – not sure what this is. This was a citation error

5. P20L5: calibrations models → calibration models corrected

6. P30L6: for new smaples. → for new samples corrected

7. P36L10: for the rediction standard error → for the prediction standard error corrected

References: There were many errors in the references section, and it would take too much time for me find and fix all of the problems. A non-exhaustive list follows, with some examples. Please redo the references section.

1. Many of the references appear to be missing journals, and then the title ends up being formatted like the journal. Examples include Cunningham et al (1976), Efron and Tibshirani (1996).

   We thank the reviewer for catching these errors. They have been corrected and all references have been reviewed.

2. The Debus et al (2018) citation has no information other than author, title, and year, and the fact that it was accepted in some journal. This is a problem because section 2.3 cites Debus et al (2018) heavily. I tried to find it by searching online but could not.

   We apologize for the error. At the time, the manuscript was only under review, but now is accepted and available online http://dx.doi.org/10.1177/0003702818804574.

3. Book and grey literature references are incorrect, e.g. Mahalanobis, P "On the generalised distance in statistics", Tibshirani (2014) References:

   - 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.
   - Solomon, P. A., Crumpler, D., Flanagan, J. B., Jayanty, R. K. M., Rickman, E. E., and McDade, C. E.: U.S. National PM2.5 Chemical Speciation Monitoring Network- CSN and IMPROVE: Description of networks, Journal of the Air & Waste Management Association, 64, 1410-1438, 10.1080/10962247.2014.956904, 2014.
   - Weakley, A. T., Takahama, S., and Dillner, A. M.: Ambient aerosol composition by infrared spectroscopy and partial least-squares in the chemical speciation network: Organic carbon with functional group identification, Aerosol Science and Technology, 50, 1096-1114, 10.1080/02786826.2016.1217389, 2016.

   We thank the reviewer for the additional references for the monitoring networks — they have now been included in the paper.

Additional comments from Quick Review Report

This paper appears to be quite good and should be given further review. My initial thought was to wonder if this was really a 'review article' - or just a very (high quality) extensive paper which describes the methods better. For example, the 'number of samples' section (3.5.2) cites only one paper, which was written by one of the co-authors. However, in the end, I thought that it has the potential to be the kind of paper I would want to read if I did want an overview of the steps taken by the authors, so I think it is OK as a review article.

In addition to a detailed description of methods, we believe it is a review in the sense that the paper synthesizes findings from past work on the topic of calibration with FT-IR spectra with collocated measurements. (The review covers mostly by the authors' work as there has not been much work published in this regard.) On the topic of carbon estimation, for instance, spectral preparation and model selection have been treated

differently in different works, and so this paper provides a broader perspective in which commonalities (and differences) in the approaches are discussed.

A few minor points:

1. The table of contents should be removed. I surveyed several other review articles in AMT and did not find any table of contents. However, this is a long article, perhaps the authors could consider adding it to a supplement.

   We have moved this to the appendix.

2. The authors should add an appendix which has definitions of all acronyms. It would help for readability in the interactive discussion. Again, this is a long article, it can be hard to find where an acronym is defined.

   We have added Table B1 which covers acronyms used in multiple sections.

3. Similarly, I recommend also defining site locations in the appendix (e.g. BYIS).

   We have included site acronyms that are used in multiple sections in Table B1.

4. Data Availability: The paper states that the data from IMPROVE and STN will be made available - does that apply to the FTIR spectra as well?

   The FT-IR spectra will be made publicly available, but at current time can be obtained by request from the authors.

Technical corrections:
'Atmosphere' is misspelled on line 5 of the abstract corrected
There is only TEXT for the copyright statement (Page 3, line 4) This will be revised upon publication.

Atmospheric particulate matter characterization by Fourier Transform Infrared spectroscopy: a review of statistical calibration strategies for carbonaceous aerosol quantification in US measurement networks

Satoshi Takahama[1], Ann M. Dillner[2], Andrew T. Weakley[2], Matteo Reggente[1], Charlotte Bürki[1], Mária Lbadaoui-Darvas[1], Bruno Debus[2], Adele Kuzmiakova[1,7], and Anthony S. Wexler[2,3,4,5,6]

[1]ENAC/IIE Swiss Federal Institute of Technology Lausanne (EPFL), Lausanne, Switzerland
[2]Air Quality Research Center, University of California Davis, Davis, CA 95616, United States
[3]Center for Health and the Environment, University of California, Davis, CA 95616, United States
[4]Mechanical and Aeronautical Engineering, University of California, Davis, CA 95616, United States
[5]Civil and Environmental Engineering, University of California, Davis, CA 95616, United States
[6]Land, Air and Water Resources, University of California, Davis, CA 95616, United States
[7]now at Stanford University

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

Abstract.

Atmospheric particulate matter (PM) is a complex mixture of many different substances, and requires a suite of instruments for chemical characterization. Fourier Transform Infrared (FT-IR) spectroscopy is a technique that can provide quantification of multiple species provided that accurate calibration models can be constructed to interpret the acquired spectra. In this capacity, FT-IR has enjoyed a long history in monitoring gas-phase constituents in the atmosphere and in stack emissions. However, application to PM poses a different set of challenges as the condensed-phase spectrum has broad, overlapping absorption peaks and contributions of scattering to the mid-infrared spectrum. Past approaches have used laboratory standards to build calibration models for prediction of inorganic substances or organic functional groups and predicting their concentration in atmospheric PM mixtures by extrapolation.

In this work, we review recent studies pursuing an alternate strategy, which is to build statistical calibration models for mid-IR spectra of PM using collocated ambient measurements. Focusing on calibrations with organic carbon (OC) and elemental carbon (EC) reported from thermal optical reflectance (TOR), this synthesis serves to consolidate our knowledge for extending FT-IR to provide TOR-equivalent OC and EC measurements to new PM samples when TOR measurements are not available. We summarize methods for model specification, calibration sample selection, and model evaluation for these substances at several sites in two US national monitoring networks: 7 sites in the Interagency Monitoring of PROtected Visual Environments (IMPROVE) network for the year 2011, and 10 sites in the Chemical Speciation Network (CSN) for the year 2013. We then describe application of the model in an operational context for the IMPROVE network for samples collected in 2013 at 6 of the same sites as 2011, and 11 additional sites. In addition to extending the evaluation to samples from a different year and different sites, we describe strategies for error anticipation due to precision and biases from the calibration model to assess

model applicability for new spectra a priori. We conclude with a discussion regarding past work and future strategies for recalibration. In addition to targeting numerical accuracy, we encourage model interpretation to facilitate understanding of the underlying structural composition related to operationally-defined quantities of TOR OC and EC from the vibrational modes in mid-IR deemed most informative for calibration. The paper is structured such that the life cycle of a statistical calibration model for FT-IR can be envisioned for any substance with IR-active vibrational modes, and more generally for instruments requiring ambient calibrations.

1 Introduction

Airborne particles are made of inorganic salts, organic compounds, mineral dust, black carbon, trace elements, and water (Seinfeld and Pandis, 2016). While regulatory limits on airborne particulate matter (PM) concentrations are set by gravimetric mass determination, analysis of chemical composition is desired as it provides insight into source contributions, facilitates evaluation of chemical simulations, and strengthens links between particle constituents and health and environmental impacts. However, the diversity of molecular constituents pose challenges for characterization as no single instrument can measure all relevant properties; an amalgam of analytical techniques are often required for comprehensive measurement (Hallquist et al., 2009; Kulkarni et al., 2011; Pratt and Prather, 2012; Nozière et al., 2015; Laskin et al., 2018). Fourier transform infrared (FT-IR) spectroscopy is one analytical technique that captures the signature of a multitude of PM constituents that give rise to feature-rich spectral patterns over the mid-infrared (mid-IR) wavelengths (Griffiths and Haseth, 2007). In the past decade, mid-IR spectra have been used for quantification of various substances in atmospheric PM, and for apportionment of organic matter (OM) into source classes including biomass burning, biogenic aerosol, fossil fuel combustion, and marine aerosol (Russell et al., 2011). The quantitative information regarding the abundance of substances in each spectrum is limited only by the calibration models that can be built for it.

In principle, the extent of frequency-dependent absorption in the mid-IR accompanying induced changes in the dipole moment of molecular bonds can be used to estimate the quantity of sample constituents in any medium (Griffiths and Haseth, 2007). Based on this principle, FT-IR has a long history in remote and ground-based measurement of chemical composition in the atmospheric vapor phase (Griffith and Jamie, 2006). For ground-based measurement, gases are measured by FT-IR in an open-path in-situ configuration (Russwurm and Childers, 2006), or via extractive sampling into a closed, multi-pass cell (Spellicy and Webb, 2006). These techniques have been used to sample urban smog (Pitts et al., 1977; Tuazon et al., 1981; Hanst et al., 1982); smog chambers (Akimoto et al., 1980; Pitts et al., 1984; Ofner, 2011), biomass burning emissions (Hurst et al., 1994; Yokelson et al., 1997; Christian et al., 2004), volcanoes (Oppenheimer and Kyle, 2008), and fugitive gases (Kirchgessner et al., 1993; Russwurm, 1999; U.S. EPA, 1998); emission fluxes (Galle et al., 1994; Griffith and Galle, 2000; Griffith et al., 2002), greenhouse gases (Shao and Griffiths, 2010; Hammer et al., 2013; Schütze et al., 2013; Hase et al., 2015);

and isotopic composition (Meier and Notholt, 1996; Flores et al., 2017). For these applications, quantitative analysis has been conducted using various regression algorithms with standard gases or synthetic calibration spectra with absolute accuracies on the order of 1–5%. Synthetic spectra for calibration are generated from a database of absorption line parameters together with simulation of pressure and Doppler broadening, and instrumental effects (Griffith, 1996; Flores et al., 2013).

1.1 Limits of conventional approaches to calibration

Analysis of FT-IR spectra of condensed-phase systems are more challenging. PM can be found in crystalline solid, amorphous solid, liquid, and semi-solid phase states (Virtanen et al., 2010; Koop et al., 2011; Li et al., 2017). Solid and liquid-phase spectra do not have the same rotational lineshapes present in the vapor phase, but inhomogeneous broadening occurs due to a multitude of local interactions of bonds within the liquid or solid environment (Turrell, 2006; Griffiths and Haseth, 2007; Kelley, 2013). Lineshapes are particularly broad in complex mixtures of atmospheric PM, since the resulting spectrum is the superposition of varying resonances for a given type of bond. FT-IR has enjoyed a long history of qualitative analysis of molecular characteristics in multicomponent PM based on visible peaks in the spectrum (e.g., Mader et al., 1952; Presto et al., 2005; Kidd et al., 2014; Chen et al., 2016a), and study of relative composition or changes to composition under controlled conditions (e.g., humidification, oxidation) has provided insight into atmospherically-relevant aerosol processes (e.g., Cziczo et al., 1997; Gibson et al., 2006; Hung et al., 2013; Zeng et al., 2013). Quantitative prediction of substances in collected PM presents a separate task, and is conventionally pursued by generating laboratory standards and relating observed features to known concentrations. This calibration approach has been predominantly used to characterize ambient and atmospherically-relevant particles collected on filters or optical disks. The bulk of past work in aerosol studies have focused on using laboratory standards to build semi-empirical calibration models for individual vibrational modes belonging to one of many functional groups present in the mixture. In this approach, the observed absorption is related to a reference measurement (typically gravimetric mass) of the compounds on the substrate. In this way, calibration of nitrate and sulfate salts (Cunningham et al., 1974; Cunningham and Johnson, 1976; Bogard et al., 1982; McClenny et al., 1985; Krost and McClenny, 1992, 1994; Pollard et al., 1990; Tsai and Kuo, 2006; Reff et al., 2007), silica dust (Foster and Walker, 1984; Weakley et al., 2014; Wei et al., 2017), and organic functional groups (Allen and Palen, 1989; Paulson et al., 1990; Pickle et al., 1990; Mylonas et al., 1991; Palen et al., 1992, 1993; Holes et al., 1997; Blando et al., 1998; Maria et al., 2002, 2003; Sax et al., 2005; Gilardoni et al., 2007; Reff et al., 2007; Coury and Dillner, 2008; Day et al., 2010; Takahama et al., 2013; Faber et al., 2017) have been studied. The organic carbon and organic aerosol mass reconstructed has typically ranged between 70–100% when compared with collocated evolved-gas analysis or mass spectrometry measurements (Russell et al., 2009; Corrigan et al., 2013), though many model uncertainties remain. One is that unmeasured, non-functionalized skeletal carbon can lead to less than full mass recovery, and the second is the estimation of the detectable fraction due to the multiplicity of carbon atoms associated with each type of functional group. (Maria et al., 2003; Takahama and Ruggeri, 2017). The challenge in this type of calibration is in the problem of extrapolating from the reference composition, which is necessarily kept simple, to that of the chemically complex PM. Spectroscopically, this difference can lead to shifts in absorption intensity or peak locations, and a general broadening of

absorption peaks on account of the same functional group appearing in many different molecules and in different condensed-phase environments.

Synthetic spectra for condensed-phase systems can be generated by mechanistic and statistical means, but are not readily available for quantitative calibration. Absolute intensities are typically even more difficult to simulate accurately for than peak frequencies (Gussoni et al., 2006). Computational models that predict vibrational motion of molecules in isolation using quantum mechanical models (Barone et al., 2012) or by harmonic approximation for larger molecules (Weymuth et al., 2012) suffer from two shortcomings: poor treatment of anharmonicity and lack of solvent effects in liquid solutions (Thomas et al., 2013). Quantum mechanical simulations can parameterize interactions with an implicitly modeled solvent through a polarizable continuum model framework (Cappelli and Biczysko, 2011), but do not adequately represent specific interactions such as hydrogen bonding (Barone et al., 2014). Microsolvation can be a better technique to describe hydrogen bonding environment but the high computational cost prevents application to large systems (Kulkarni et al., 2009). Gaussian dispersian analysis has provided accurate spectrum reconstruction in pure liquids (water-ethanol mixtures) from their calculated dielectric functions (MacDonald and Bureau, 2003), but has not been applied to more complex systems. Molecular dynamics (MD) provides a general framework for addressing interactions with the solvent, large-amplitude motions in flexible molecules, and anharmonicities (Ishiyama and Morita, 2011; Ivanov et al., 2013). Electronic structure calculations relevant for predicting vibrational spectra can be incorporated by ab initio MD (Car and Parrinello, 1985; Marx, 2009; Thomas et al., 2013), and path integral MD methods such as centroid or ring polymer MD (Witt et al., 2009; Ceriotti et al., 2016) that additionally considers nuclear quantum effects (at higher computational cost). Ab initio MD is widely used for the simulating the spectra of water and a range small organic and biological molecules in isolation (Silvestrelli et al., 1997; Aida and Dupuis, 2003; Gaigeot et al., 2007; Gaigeot, 2008; Thomas et al., 2013; Fischer et al., 2016) Such calculations generally reproduce the shape of the spectrum well with respect to experimental ones at very high dilution, although C-H stretching peaks are known to be shifted towards higher wavenumbers due to the lack of improper hydrogen bonding in vacuum simulations (Thomas et al., 2013). Bulk liquid phase simulations are limited to a few tens of molecules (few hundreds of atoms), and have been performed for liquids, including methanol (Thomas et al., 2013), water (Silvestrelli et al., 1997), and aqueous solutions of biomolecules (Gaigeot and Sprik, 2003). These simulations reproduce peak positions and relative intensities sufficiently well when compared to experimental spectra, albeit with lower accuracy in peak position at wavenumbers higher than 2000 cm^{-1}. These methods have also been shown to reproduce main features of vibrational spectra in solid (crystalline ice and naphthalene) systems (Bernasconi et al., 1998; Putrino and Parrinello, 2002; Pagliai et al., 2008; Rossi et al., 2014b). Nuclear quantum effects not explicitly accounted for by ab initio calculations become more important for hydrogen-containing systems, and have been investigated in liquid water and methane for vibrational spectra simulation (Rossi et al., 2014a, b; Medders and Paesani, 2015; Marsalek and Markland, 2017). A recent approach improves upon the accuracy and speed of ab initio MD by combining a dipole moment model (Gastegger et al., 2017) and potentials (Behler and Parrinello, 2007) derived from machine learning. Trained on only several hundred reference electronic structure calculations, spectra of several alkanes and small peptides were simulated with accuracy reflecting improved treatment of anharmonicities and proton transfer, with reductions in computational cost by three orders of magnitude (Gastegger et al., 2017). However, this machine-learned method still inherits some common limitations of ab initio

calculations upon which models are trained. One example is the apparent blue shift of the C-H stretching peak, likely due to an insufficient treatment of improper hydrogen bonding or the deficiency of the electron exchange functional (Thomas et al., 2013). While such methods may be useful in aiding interpretation of environmental spectra (Kubicki and Mueller, 2010; Pedone et al., 2010), they are not yet mature for reproducing spectra of suitable quality for quantitative calibration or (white-box) inverse modeling.

Early applications of artificial intelligence to mid-IR spectra interpretation also included efforts to generate synthetic spectra of individual compounds. Mid-IR spectra of new compounds were simulated from neural networks trained on three-dimensional molecular descriptors (radial distribution functions) paired with corresponding mid-IR spectra, matched by similarity (nearest neighbor) search in a structural database, or generated from substructure/spectral correlation databases (Dubois et al., 1990; Weigel and Herges, 1996; Baumann and Clerc, 1997; Schuur and Gasteiger, 1997; Selzer et al., 2000; Yao et al., 2001; Gasteiger, 2006). Drawing upon internal or commercial libraries (Barth, 1993), predictions were made for compounds in the condensed phase with a diverse set of substructures including including methanol, amino acids, ring-structured acids, and substituted benzene derivatives. Many structural features including peak location, relative peak heights, and peak widths were reproduced, provided that relevant training samples were available in the library. Much of the work was motivated by pattern matching and classification of spectra for unknown samples (Robb and Munk, 1990; Novic and Zupan, 1995), and automated band assignment and identification of the underlying fragments typically performed by trained spectroscopists (Sasaki et al., 1968; Gribov and Elyashberg, 1970; Christie and Munk, 1988; Munk, 1998; Hemmer, 2007; Elyashberg et al., 2009). This approach has been able to generate spectra for more complex molecules than mechanistic modeling relying on ab initio calculations. However, the extent of evaluation has been limited; extension to multicomponent mixtures and usefulness for quantitative calibration is currently not known. While these research fields remain an active part of cheminformatics, we propose another approach for calibration model development that can be used for atmospheric PM analysis.

1.2 Use of collocated measurements

As an alternative to laboratory-generated mixtures and simulated spectra, collocated measurements of substances for which there are IR-active vibrational modes can be used as reference values for calibration (also referred to as "in-situ" calibration). This data-driven approach permits the complexity of atmospheric PM spectra with overlapping absorbances from both analytes and interferences to be included in a calibration model. For instance, Allen et al. (1994) demonstrated the use of collocated ammonium sulfate measurements by ion chromatography to quantify the abundance of this substance from FT-IR spectra, though some uncertainties arose from the time resolution between the sampling instruments.

The benefit of building data-driven calibration models to reproduce concentrations reported by available measurements is twofold. One is to provide equivalent measurements when the reference measurements are expensive or difficult to obtain. For example, FT-IR spectra can be acquired rapidly, non-destructively, and at low cost from from Polytetrafluoroethylene (PTFE) filters commonly used for gravimetric mass analysis in compliance monitoring and health studies. That vibrational spectra contain many signatures of chemical constituents of PM (which also gives rise to challenges in spectroscopic interpretation) provides the basis for quantitative calibration of a multitude of substances. This capability for multi-analyte analysis is benefi-

cial when a single filter may be relied upon during short-term campaigns, or in network sites for which installation of the full suite of instruments is prohibitive. The second benefit is the ability to gain a better understanding of atmospheric constituents measured by other techniques by associating them with important vibrational modes structural elements of molecules identified in the FT-IR calibration model. Such an application can be enlightening for studying aggregated metrics such as carbon content, or functional group composition in atmospheric PM quantified by techniques requiring more sample mass and user labor: ultraviolet-visible spectrometry or nuclear magnetic resonance spectroscopy (Decesari et al., 2003; Ranney and Ziemann, 2016).

In this paper, we demonstrate an extensive application of this approach in the statistical calibration of FT-IR spectra to collocated measurements of carbonaceous aerosol content — organic carbon (OC) and elemental carbon (EC) — characterized by a particular type of evolved gas analysis (EGA). EGA includes thermal optical reflectance (TOR) and thermal optical transmittance (TOT), which apportions total carbon into OC and EC fractions according to different criteria applied to the changing optical properties of the filter under stepwise heating (Chow et al., 2007a). EGA OC and EC are widely-measured in monitoring networks (Chow et al., 2007a; Brown et al., 2017), with historical significance in regulatory monitoring, source apportionment, and epidemiological studies. While EC is formally defined as sp^2-bonded carbon bonded only to other carbon atoms, EC measured by EGA is an operationally-defined quantity which is likely associated with low-volatility organic compounds (Chow et al., 2004; Petzold et al., 2013; Lack et al., 2014). EGA OC comprises a larger fraction of the total carbon and therefore is less influenced by pyrolysis artifacts that affects quantification of EGA EC. In addition to OC estimates independently constructed from laboratory calibrations of functional groups, prediction of EGA OC and EC from FT-IR spectra will provide values for which strong precedent in atmospheric studies exist. Thus, use of collocated measurements complement conventional approaches in expanding the capabilities of FT-IR spectroscopy to extract useful information contained in vibrational spectra.

We review the current state-of-the art for quantitative prediction of OC and EC as reported by TOR using FT-IR in selected sites of the Interagency Monitoring of PROtected Visual Environments (IMPROVE) monitoring network (Malm and Hand, 2007; Solomon et al., 2014) and the Chemical Speciation Network (CSN) (Solomon et al., 2014). This work is placed within the context of overseeing the life cycle of a statistical calibration model more generally; reporting further developments in anticipating errors due to precision and bias in new samples, and describing a roadmap for future work. While partial least squares (PLS) regression and its variants figure heavily in the calibration approach taken thus far, related developments in the fields of machine learning, chemometrics, and statistical process monitoring are mentioned to indicate the range of possibilities yet available to overcome future challenges in interpreting complex the mid-IR spectra of PM. We expect that many concepts described here will also be relevant for the emerging field of statistical calibration and deployment of measurements in a broader environmental and atmospheric context (e.g., Cross et al., 2017; Kim et al., 2018; Zimmerman et al., 2018). In the following sections, we describe the experimental methods for collecting data (Section 2), the calibration process (Section 3), assessing suitability of existing models for new samples (Section 4.1), and maintaining calibration models (Section 4.2). Finally, we conclude with a summary and outlook (Section 5). A table of contents is included in Section A and list of recurring acronymns in Section B.

2 Background

First, we review the basic principles of FT-IR and how the measured absorbances can be related to underlying constituents, including carbonaceous species (Section 2.1). We then describe the samples used for calibration and evaluation (Section 2.2). We then conclude the section with discussion regarding quality assurance and quality control (QA/QC) of the FT-IR hardware performance (Section 2.3). Under the assumption that these hardware QA/QC criteria are met, we dedicate the remainder of the paper outlining model evaluation on the assumption that the performance in prediction can be attributed to differences in sample composition.

2.1 Fourier transform-infrared spectroscopy

In this section, we cover the background necessary to understand FT-IR spectroscopy in the analysis of PM collected onto PTFE filter media, which is optically thin and permits an absorbance spectrum to be obtained by transmission without additional sample preparation (McClenny et al., 1985; Maria et al., 2003). The wavelengths of IR are longer than visible light (400–800 nm) and FT-IR refers to a non-dispersive analytical technique probing the mid-IR, which is radiation from 2,500 nm to 25,000 nm or in the vibrational frequency units used by spectroscopists, wavenumbers, 4000 to 400 cm^{-1}. Molecular bonds absorb mid-IR at characteristic frequencies of their vibrational modes when interactions between electric dipole and electric field induce transitions among vibrational energy states (Steele, 2006; Griffiths and Haseth, 2007). Based on this principle, the spectrum obtained by FT-IR represents the underlying composition of organic and inorganic functional groups containing molecular bonds with a dipole moment.

In transmission-mode analysis where the IR beam is directed through the sample, absorbance (A) can be obtained by ratioing the measured extinction of radiation through the sample (I) by a reference value (I_0), also called the "background", and taking the negative value of their decadic logarithm (first relation of eq. 1).

$$A(\tilde{\nu}) = -\log_{10}\left[\frac{I(\tilde{\nu})}{I_0(\tilde{\nu})}\right] = \epsilon(\tilde{\nu})n^{(a)} \tag{1}$$

The sample is the PTFE filter (with or without PM) and the background is taken as the empty sample compartment. The quality of the absorbance spectrum depends on how accurately the background reflects the conditions of the sample scan, and the background is therefore acquired regularly as discussed in Section 2.3.

When absorption is the dominant mode of extinction, the measured absorbance (A) is proportional to the areal density of molecules ($n^{(a)}$) in the beam in the sample (eq. 1) (Duyckaerts, 1959; Kortüm, 1969; Nordlund, 2011). ϵ is the proportionality constant and is called the molar absorption coefficient. Although scattering off of surfaces present in the sample can generate a significant contribution to the absorbance spectrum, its effects can be modeled as a sum of incremental absorbances by a linear calibration model, or minimized through spectral pre-processing procedures (baseline correction) as discussed in Section 3.3.1.

A composite metric of PM such as carbon content presumably results from contributions by a myriad of substances. The abundances of these underlying molecules concurrently give rise to the apparent mass of carbon (m_{C}) (eq. 2) measured by

evolved gas analysis and the absorbance spectrum (A) (eq. 3) measured by FT-IR (Ottaway et al., 2012):

$$m_{\mathrm{C}}^{(a)} = 12.01 \cdot \sum_{k} f_{\mathrm{C},k} n_k^{(a)} \tag{2}$$

$$A(\tilde{\nu}) = \sum_{k} \epsilon_k(\tilde{\nu})\, n_k^{(a)} + \sum_{k'} \epsilon_{k'}(\tilde{\nu})\, n_{k'}^{(a)} + \{\ldots\} \ . \tag{3}$$

$f_{\mathrm{C},k}$ denotes the number of (organic or elemental) carbon in molecule k, and 12.01 is the atomic mass of carbon. Non-carbonaceous substances (e.g., inorganic compounds) that give rise to additional (possibly interfering) absorbance are indexed by k'. The superscript "(a)" denotes an area-normalized quantity. "$\{\ldots\}$" indicates contributions from instrumental noise, ambient background, and additional factors such as scattering. Using TOR measurements from collocated quartz fiber filters, our objective is to develop a calibration model for estimating the abundance of carbonaceous material ($m_{\mathrm{C}}^{(a)}$) in the PTFE sample that may have led to the observed pattern of mid-IR absorbances ($A(\tilde{\nu})$). A common approach is to explore the relationship between response and absorbance spectra through a class of models which take on a multivariate, linear form (Griffiths and Haseth, 2007):

$$m_{\mathrm{C},i}^{(a)} = \sum_{j} b_j A_i(\tilde{\nu}_j) + e_i \ . \tag{4}$$

The set of wavelength-dependent regression coefficients b_j comprise a vector operator that effectively extracts the necessary information from the spectrum for calibration. These coefficients (b_js) presumably represent a weighted combination of coefficients expressed in eqs. 2 and 3 (also correcting for non-carbonaceous interferences). The remaining term, e_i, characterizes the model residual (in regression fitting) or prediction error (in application to new samples). The relationship with underlying substances (k) that comprise OC and EC is implicit, though some efforts to interpret these constituents have been made through examination of latent (or hidden) variables obtained from the calibration model (discussed in Section 3.4).

Using complex, operationally-defined TOR measurements as reference for calibration, some caution in interpretation and application is warranted. For instance, these coefficients may not necessarily capture the true relationship expressed by eqs. 2 and 3, but rather rely on correlated rather than causal variables for quantification. Particles and the PTFE substrate itself can confer a large scattering contribution to the extinction spectrum (eq. 1), and additional sample matrix interactions among analytes may challenge assumptions regarding the linear relationship (eq. 3) underlying the model for quantification (eq. 4) (Geladi and Kowalski, 1986). Furthermore, the relationship between spectra and concentrations embodied by $\boldsymbol{b}$ is specific to the the chemical composition of PM at the geographic location and sampling artifacts due to composition and sample handling protocols of the calibration samples.To address these concerns, extensive evaluation regarding model performance in various extrapolation contexts are necessary to investigate the limits of our calibration models, and methods for anticipating prediction errors provide some guidance on their general applicability in new domains. Regression coefficients and underlying model parameters are inspected to determine important vibrational modes that provide insight into the infrared absorption bands that drive the predictive capability of our regression models.

2.2 Sample collection (IMPROVE and CSN)

The IMPROVE network consists of approximately 170 sites in rural and pristine locations in the United States primarily National Parks and Wilderness Areas (Malm and Hand, 2007). Data from the IMPROVE network is used to monitor trends in particulate matter concentrations and visibility. IMPROVE collects ambient samples midnight to midnight every third day by pulling air at 22.8 liters per minute through filters. Polytetrafluoroethylene (PTFE, 25 mm, Pall Corp.) or more commonly referred to as Teflon filters are routinely used for gravimetric, elemental, and light absorption measurements and are used in this work for FT-IR analysis. Quartz filters are used for thermal optical reflectance (TOR) measurements to obtain organic and elemental carbon. Nylon filters are used to measure inorganic ions, primarily sulfate and nitrate.

The CSN consists of about 140 sites located in urban and suburban area and the data is used to evaluate trends and sources of particulate matter (Solomon et al., 2014). Ambient samples are collected in the CSN on midnight to midnight schedule one every third or one every sixth day. Quartz filters for TOR analysis are collected with a flowrate of 22.8 lpm. PTFE filters (Whatman $PM_{2.5}$ membranes, 47 mm, used through late 2015; MTL filters (Measurement Technology Laboratories, 47 mm) have been used there after) and nylon filters are collected at a flowrate of 6.7 lpm ~~flowrate~~. All sites in CSN have used TOR ~~analysis~~ for carbon analysis since 2010.

PTFE filters are used for gravimetric analysis on account of its low vapor absorption (especially water) and standardization in compliance monitoring, while quartz fiber filters are separately collected on account of its thermal stability (Chow, 1995; Chow et al., 2007b; Malm et al., 2011; Solomon et al., 2014; Chow et al., 2015). TOR analysis consists of heating a portion of the quartz filter with the IMPROVE_A temperature ramp and measuring the evolved carbon (Chow et al., 2007a). The initial heating is performed with an inert environment and the material that is removed is ascribed to organic carbon (OC). Oxygen is added at the higher temperatures and the measured material is ascribed to elemental carbon (EC). Charring of ambient particulate carbon is corrected using a laser that reflects off the surface of the sample (hence reflectance) (Chow et al., 1993). The evolved carbon is converted to methane and measured with a flame ionization detector. Organic carbon data is corrected for gas-phase adsorption using a monthly median blank value specific to each network (Dillner, 2018).

For this work, we examine a subset of these sites in which PTFE filters were analyzed for FT-IR spectra (Figure 1). For model building and evaluation (Section 3), we use 7 sites consisting of 794 samples for IMPROVE in 2011, and 10 sites consisting of 1035 samples for CSN in 2013. Two sites in 2011 IMPROVE are samplers collocated at the same urban location in Phoenix, AZ, and one site (Sac and Fox) that was discontinued mid-year. Additional IMPROVE samples were analyzed by FT-IR during sample year 2013, which included 6 of the same sites and 11 additional sites. This data set is used for evaluation of the operational phase of the model (Section 4).

Given the different sampling protocols that result in different spectroscopic interferences from PTFE (due to different filter types) and range of mass loadings (due to flowrates), and difference in expected chemical composition (due to site types), calibrations for the CSN and IMPROVE networks have been developed separately (Weakley et al., 2016). Advantages of building such specialized models in favor of larger, all-inclusive models are discussed in Section 3.5. Therefore, TOR-equivalent carbon predictions for 2011 and 2013 IMPROVE samples discussed for this paper are made with a calibration model using a subset of

samples from 2011 IMPROVE, and TOR predictions for 2013 CSN samples are made with a calibration model using a subset of samples from 2013 CSN. One exception is a special model constructed to illustrate how new samples can improve model prediction (Section 4.2); a subset of samples from two sites — Fresno, CA (FRES) and Baengnyeong Island, S. Korea (BYIS) — in 2013 IMPROVE are used to make predictions for the remaining samples at those sites. In all cases, analytical figures of merit for model evaluation are calculated for samples that are not used in calibration.

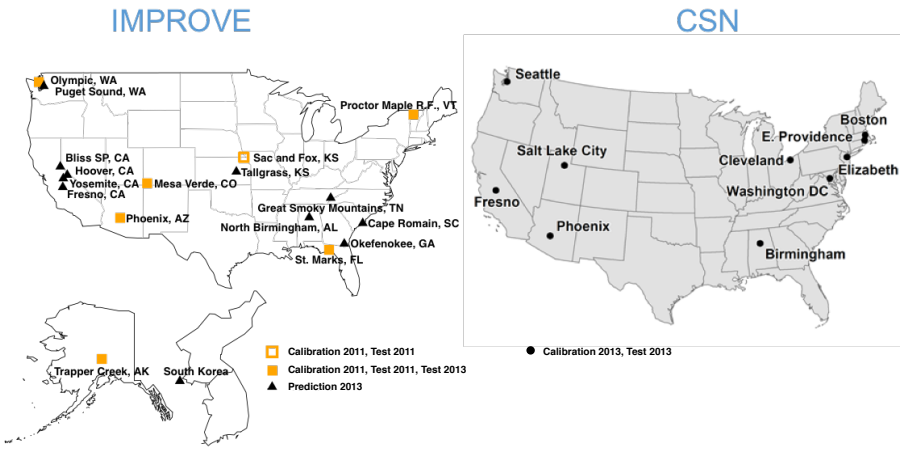

Figure 1. Map of IMPROVE and CSN sites used for this work. The Sac and Fox, KS, IMPROVE site was only operational for the first half of 2011. Samples from Fresno, CA, South Korea were additionally used for a separate calibration.

2.3 Laboratory operations and quality control of analysis

IMPROVE and CSN PTFE sample and blank filters are analyzed without pretreatment on either Tensor 27 or Tensor II FT-IR instruments (Bruker Optics, Billerica, MA) equipped with a liquid nitrogen-cooled detector. Filters are placed in a small, custom-built sample chamber which reliably places each filter the same distance from the source. IR-active water vapor and CO_2 are purged from the sample compartment and instrument optics to minimize absorption bands of gas phase compounds in the aerosol spectra. Samples are measured in transmission mode and absorbance spectra, which are used for calibration and prediction, are calculated using the most recent empty chamber spectrum as a reference (collected hourly). The total measurement time for one filter is 5 minutes. Additional details on the FT-IR analysis are described by Ruthenburg et al. (2014) and Debus et al. (2018).

Daily and weekly quality control checks are performed to monitor the comparability, precision and stability of the FT-IR instruments. Duplicate spectra are collected every fifty filters (once or twice per day) per instrument in order to evaluate measurement precision. Measured precision values are low and smaller than the 95th percentile of the standard deviation of the blanks for both TOR OC and EC indicating that instrument error has a relatively minor influence on the prediction of TOR OC and EC and is smaller than the variability observed between PTFE filters. Quality control filters — blank filters and ambient samples — are analyzed weekly to monitor instrument stability. Debus et al. (2018) conclude that predictions of TOR

OC and EC remain relatively stable over a two and a half year period based on analyses of quality control filters, and that observed changes are small. These data enable us to track instrumental changes that will require re-calibration (Section 4.2). A subset of ambient filters are analyzed on all FT-IR instruments to evaluate spectral dissimilarities and differences in prediction. These samples show that differences in spectral response between instruments are small and due mainly to variability in PTFE. In addition, these samples indicate that careful control of laboratory conditions and detector temperature, sample position, relative humidity (RH) and CO_2 levels in the FT-IR instrument enables instrument-agnostic calibrations that predict accurate concentrations independent of the instrument on which a spectrum is collected. The quality control data show that the TOR OC and EC measurements obtained from multiple FT-IR instruments in one laboratory are precise, stable (over the 2.5 year period evaluated) and agnostic to instrument used for analysis (Debus et al., 2018).

3 Model building, evaluation, and interpretation

In this section, we describe the model building process for quantitative calibration. The relationship betwen spectra and reference values to be exploited for prediction can be discovered using any number of algorithms, method of spectra pretreatment, and the *calibration set* of samples to be used for model training and validation. As the best choices for each of these categories are not known a priori, the typical strategy is to generate a large set of candidate models and select one that scores well across a suite of performance criteria against a *test set* of samples reserved for independent evaluation. The process of building and evaluating a model conceptualized in the framework of statistical process control is depicted in Figure 2. In the first stage, various pathways to model construction are evaluated, and expectations for model performance are determined. The second stage involves continued application and monitoring of model suitability for new samples (*prediction set*), which is discussed in Section 4.1. Where applicable, the sample type in each data set should include several types of samples. For instance, the calibration set can include blank samples in which analyte (but not necessarily interferent) concentrations are absent. Test and prediction set samples can include both analytical and field blank samples. Collocated measurements can be used for providing replicates for calibration, or used as separate evaluation of precision. Immediately below, we describe the procedure for model specification, algorithms for parameter estimation, and model selection in Section 3.1. Methods for spectra processing are described in Section 3.3, and sample selection in Section 3.5. In each section, the broader concept will be introduced and then its application to TOR will be reviewed.

3.1 Model estimation

Many algorithms in the domain of statistical learning, machine learning, and chemometrics have demonstrated utility in building calibration models with spectra measurements: neural networks (Long et al., 1990; Walczak and Massart, 2000), Gaussian process regression (Chen et al., 2007), support vector regression (Thissen et al., 2004; Balabin and Smirnov, 2011), principal components regression (Hasegawa, 2006), ridge regression (Hoerl and Kennard, 1970; Tikhonov and Arsenin, 1977; Kalivas, 2012), wavelet regression (Brown et al., 2001; Zhao et al., 2012), functional regression (Saeys et al., 2008), ~~partial least squares~~ PLS (Rosipal and Krämer, 2006); among others. There is no lack of algorithms for supervised learning with continuous re-

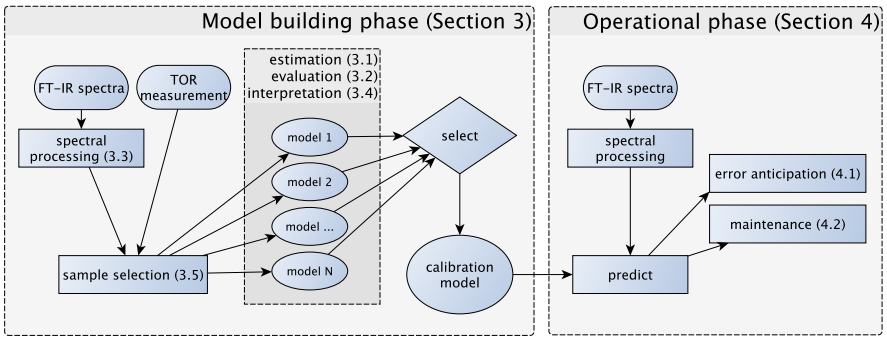

Figure 2. Diagram of the model building, evaluation, and monitoring process. Sections and subsections covering the illustrated topics are denoted in parentheses. Note that the any of the calibrations $\{1, 2, \ldots, N\}$ can be a multilevel model (Section 3.5.3) consisting of an ensemble of models.

sponse variables that can potentially be adapted for such an application (Hastie et al., 2009). Each of these techniques map relationships between spectral features and reference concentrations using different similarity measures, manifolds, and projections; largely in metric spaces where the notion of distances among real-valued data points are well-defined (e.g., Zezula et al., 2006; Russolillo, 2012). The best mathematical representation for any new data set is difficult to ascertain a priori, but models can be compared by their fundamental assumptions and their formulation: e.g., linear or non-linear in form; globally parametric, locally parametric, or distribution free (random forest, nearest neighbor); feature transformations; objective function and constraints; and expected residual distributions. Approaches that incorporate randomized sampling can return slightly different numerical results, but reproducibility of any particular result can be ensured by providing seed values for the pseudo-random number generator. A typical procedure for model development is to select candidate methods that have enjoyed success in similar applications and empirically investigate which techniques provide meaningful performance and interpretability for the current task, after which implementation measures are then pursued (Kuhn and Johnson, 2013). In lieu of selecting a single model, ensemble learning and Bayesian model averaging approaches combine predictions from multiple models (Murphy, 2012).

For FT-IR calibration targeting prediction of TOR-equivalent concentrations, we focus on finding solutions to the linear model introduced in Section 2.1. Letting $\boldsymbol{y} = [m_{\mathrm{C},i}/a]$, $\boldsymbol{X} = [A_i(\tilde{\nu}_j)]$, $\boldsymbol{b} = [b_i]$, and $\boldsymbol{e} = [e_i]$, we re-express eq. 4 in array notation to facilitate further discussions of linear operations:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e} \,. \tag{5}$$

Equation 5 is an ill-posed inverse problem; therefore, it is desirable to introduce some form of regularization (method of introducing additional information or assumptions) to find suitable candidates for $\boldsymbol{b}$ (Zhou et al., 2005; Friedman et al., 2010; Takahama et al., 2016). In this paper, we summarize the application of ~~partial least squares (PLS)~~ PLS (Wold, 1966; Wold et al., 2001) for obtaining solutions to this equation, with which good results have been obtained for our application and FT-IR spectra more generally (Hasegawa, 2006; Griffiths and Haseth, 2007). This technique has been a classic workhorse of chemometrics

for many decades and is particularly well suited for characteristics of FT-IR analysis, for which data are collinear (neighboring absorbances are often related to one another) and high-dimensional (more variables than measurements in many scenarios). These issues are addressed by projection of spectra onto an orthogonal bases of latent variables (LVs) that take a combination of spectral features, and regularization by LV selection (Andries and Kalivas, 2013). Furthermore, PLS is agnostic with respect to assumption of residual structure (e.g., normality) for obtaining $\boldsymbol{b}$, which circumvents the need to explicitly account for covariance or error distribution models to characterize the residuals (Aitken, 1936; Nelder and Wedderburn, 1972; Kariya and Kurata, 2004). PLS is also used as a preliminary dimension reduction technique prior to application of non-linear methods (Walczak and Wegscheider, 1993) ~~(an approach for outlier detection is described in Section 4.1.2)~~. Therefore, it is sensible that PLS should be selected as a canonical approach for solving eq. 5.

Mathematically, classical PLS represents a bilinear decomposition of a multivariate model in which both $\boldsymbol{X}$ and $\boldsymbol{y}$ are projected onto basis sets ("loadings") $\boldsymbol{P}$ and $\boldsymbol{q}$, respectively (Wold et al., 1983, 1984; Geladi and Kowalski, 1986; Mevik and Wehrens, 2007):

$$\begin{aligned} \boldsymbol{X} &= \boldsymbol{T}\boldsymbol{P}^T + \boldsymbol{E}_X \\ \boldsymbol{y} &= \boldsymbol{T}\boldsymbol{q}^T + \boldsymbol{e}\ . \end{aligned} \tag{6}$$

$\boldsymbol{T}$ is the orthogonal score matrix and $\boldsymbol{E}_X$ denotes the residuals in the reconstruction of the spectra matrix. Common solution methods search for a set of loading weight vectors (represented in a column matrix $\boldsymbol{W}$) such that covariance of scores ($\boldsymbol{T}$) with respect to the response variable ($\boldsymbol{y}$) is maximized. The weight matrix can be viewed as a linear operator that changes the basis between the feature space and FT-IR measurement space. These weights and their relationship to the score matrix and regression vector are expressed below:

$$\begin{aligned} \boldsymbol{R} &= \boldsymbol{W}\left(\boldsymbol{P}^T\boldsymbol{W}\right)^{-1} \\ \boldsymbol{T} &= \boldsymbol{X}\boldsymbol{R} \\ \boldsymbol{b} &= \boldsymbol{R}\boldsymbol{q}^T\ . \end{aligned} \tag{7}$$

For univariate $\boldsymbol{y}$ as written in eq. 5, a number of commonly used algorithms — Nonlinear Iterative PArtial Least Squares (NIPALS; Wold et al., 1983), SIMPLS (deJong, 1993), kernel PLS (with linear kernel; Lindgren et al., 1993) — can be used to arrive at the same solution (while varying in numerical efficiency). Kernel PLS can be further extended into modeling non-linear interactions by projecting the spectra onto a high-dimensional space and applying linear algebraic operations akin to classical PLS, with comparative performance to support vector regression and other commonly-used non-linear modeling approaches (Rosipal and Krämer, 2006). However, likely due to the linear nature of the underlying relationship (eq. 4), linear PLS has typically performed better than non-linear algorithms for FT-IR calibration (Griffiths and Haseth, 2007). In addition, the linearity of classical PLS regression has yielded more interpretable models than non-linear ones (Luinge et al., 1995). Therefore, past applications of PLS to FT-IR calibration of atmospheric aerosol constituents has focused on its linear variants and will be the focus of this paper.

An optimal number of LVs must be selected to arrive at the best predictive model. A larger number of LVs is increasingly able to capture the variations in the spectra, leading to reduction in model bias. Some of the finer variations in the spectra are

not part of the analyte signal which we wish to model; including LVs that model these terms lead to increased variance in its predictions. A universal problem in statistical modeling is to find a method for characterizing model bias and variance such that one with the lowest apparent error can be chosen. There is no shortage of methods devised to capture this bias-variance tradeoff and their implications for model selection continue to be an active area of development (Hastie et al., 2009). With no immediate consensus on the single best approach for all cases, the approach often taken is to select and use one based on prior experience until found to be inadequate (as with model specification).

One class of methods characterize the bias and variance using the information obtained from fitting of the data. For instance, Akaike Information Criterion (AIC; Akaike, 1974) and Bayesian Information Criterion (BIC; Schwarz, 1978) consider the balance between model fidelity (fitting error, which monotonically decreases with number of parameters) with penalties incurred for increasing model complexity (which serves as a form of regularization). The fitting error may be characterized by residual sum of squares or maximum likelihood estimate (e.g., Li et al., 2002), and the penalty may be a scaled form of the number of parameters or norms of the regression coefficient vector. An effective degrees of freedom (EDF) or generalized EDF parameter aims to characterize the resolvable dimensionality as apparent from the model fit to data (Tibshirani, 2014), though the EDF may not always correspond to desired model complexity (Krämer and Sugiyama, 2011; Janson et al., 2015).

Another class of methods relies on assessment of the bias and variance contributions implicitly present in prediction errors, which are obtained by application of regression coefficients estimated using a training data set and evaluated against a separate set of ("validation") data withheld from model construction to fix its parameters. To maximize the data available for both training and validation, modern statistical algorithms such as cross validation (Mosteller and Tukey, 1968; Stone, 1974; Geisser, 1975) and the bootstrap method (Efron and Tibshirani, 1997) allows use of the same samples for both training and validation, which comprise what we collectively refer to as the calibration set. The essential principle is to partition the same calibration set multiple times such that the model is trained and then validated on different samples over a repeated number of trials. In this way, a distribution of performance metrics for models containing different subsets of the data can be aggregated to determine a suitable estimate of a parameter (number of LVs). The number and arrangement of partitions vary by method, with cross-validation using each sample exactly once for validation and bootstrap resamples with replacement. Both have reported usable results (Molinaro et al., 2005; Arlot and Celisse, 2010). For increasingly smaller number of samples, Leave-One-Out (LOO) CV or bootstrap may be favored as it reserves a larger number of samples to train each model, though it is generally appreciated that LOO leads to suboptimal estimates of prediction error (Hastie et al., 2009). Evaluation metrics are calculated on samples which have not been involved in the model-building process (Esbensen and Geladi, 2010). Examples of metrics include the minimum root-mean-square error of cross validation (RMSECV) (one of the most widely used metrics; Gowen et al., 2011), one standard deviation above RMSECV (Hastie et al., 2009), Wold's R criterion (Wold, 1978), coefficient of determination (R^2), randomization p-value (van der Voet, 1994; Wiklund et al., 2007), among others. A suite of these metrics can also be considered simultaneously (Zhao et al., 2015). The final model is obtained by refitting the model to all of the available samples in the calibration set and using the number of parameters selected in the CV process. Other strategies and general discussions on the topic of performance metrics and statistical sampling are covered in many textbooks (e.g., Bishop, 2009; Hastie et al., 2009; Kuhn and Johnson, 2013).

Past work on TOR and FT-IR measurements have used V-fold CV, with Dillner and Takahama (2015a; 2015b) using minimum RMSECV and Weakley et al. (2016) using Wold's R criterion for performance evaluation. In V-fold CV, the data is partitioned into V groups, and V-1 subsets are used to train a model to be evaluated on the remaining subset (repeated for V arrangements). Dillner and Takahama (2015a) found that V=2, 5, and 10 selected different number of LVs but led to similar overall performance. To keep the solution deterministic (i.e., no random sampling) and representative (i.e., the composition of training sets and validation sets are representative of the overall calibration sets across permutations), samples in the calibration set are ordered according to a strategy amenable for stratification. For instance, samples are arranged by sampling site and date (used as a surrogate for source emissions, atmospheric processing, and composition, which often vary by geography and season), or with respect to increasing target analyte concentration, and samples separated by interval V are used to create each partition in a method referred to as Venetian blinds (also referred to as interleaved or striped) CV. An illustration of RMSECV compared to the fitting errors represented by the root-mean-square error calibration (RMSEC) for TOR OC is shown in Figure 3. Other strategies for arranging CV include maximizing differences among samples in each fold to reduce chances of overfitting (Kuhn and Johnson, 2013) but has not been explored in this application.

Even with specification of model and approach for parameter selection fixed, spectral processing and sample selection can lead to differences in overall model performance. We first discuss how different models can be generated from the same set of samples according to these decisions before proceeding to protocols for model evaluation using the *test set* reserved for independent assessment (Section 3.2). The test set is used to compare the merits of models built in different ways, and establish control limits for the operational phase (Section 4).

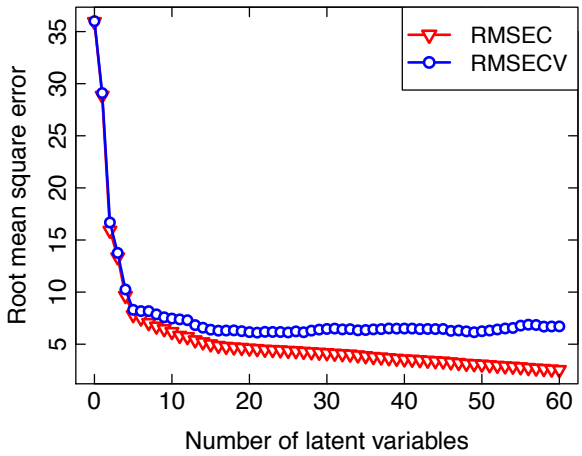

Figure 3. Illustration of RMSEC, which represents the fitting errors, and RMSECV, which represents the prediction error, calculated for TOR OC using the same calibration set. 10-fold Venetian blinds CV was used for this calculation.

3.2 Model evaluation

Statistical models can be evaluated using many of the same techniques also used by mechanistic models (Olivieri, 2015; Seinfeld and Pandis, 2016). In this section, we describe methods for evaluating overall performance (Section 3.2.1) and occurrence of systematic errors (Section 3.2.2).

3.2.1 Overall performance

Predictions for a set of selected models for 2011 IMPROVE and 2013 CSN ~~data sets~~ are shown in Figure 4. Details of sample selection for calibration are provided in Section 3.5), but here we present results for the "base case" models which contain representations of all sites and season for each network. There are many aspects of each model which we wish to evaluate by comparing predictions against known reference values. These aspects include the bias and magnitude of dispersion, but also our capability to distinguish ambient samples from blank samples at the low end of observed concentrations. Metrics which capture these effects can effectively be derived from the term $\boldsymbol{e}$ in the multivariate regression equation (eq. 5) when predictions and observations are compared in the test set spectra. $\boldsymbol{e}$ is referred to as the residual when describing deviations from observations in fitted values, and prediction error when describing deviations from observed values when the model is used for prediction in new samples. However, by convention we often resort to the negative of the residual such that deviation in prediction is calculated with respect to the observation, rather than the other way around. Example distributions for residuals and prediction errors for TOR OC in 2011 IMPROVE are shown in Figure 5.

While the use of the minimum root-mean-square error (RMSE) is pervasive in chemometrics and machine learning as a formal parameter tuning or model selection criterion, another family of metrics are more commonly used in the air quality community (Table 1). For instance, the mean bias and mean error and their normalized quantities are often used for model-measurement evaluation of mechanistic (chemical transport) models (Seinfeld and Pandis, 2016). R^2 is commonly used in inter-comparisons of analytical techniques. Many of the statistical estimators in Table 1 converge to a known distribution from which confidence intervals can be calculated; or otherwise estimated numerically (e.g., by bootstrap). In addition to conventional metrics, alternatives drawing upon robust statistics (Huber and Ronchetti, 2009) are also useful when undue influence from a few extreme values may to lead to misrepresentation of overall model performance (Barnett and Lewis, 1994). For instance, the mean bias is replaced by the median bias, and mean absolute error is replaced by median absolute deviation. Even if a robust estimator is unbiased, it may not have the same variance properties as its non-robust counterpart (Venables and Ripley, 2003); therefore, comparison against a reference distribution for statistical inference may be less straightforward.

For TOR-equivalent values predicted by FT-IR, the median bias and errors have been typically preferred for characterizing overall model performance, together with R^2 and the minimum detection limit (MDL). Mean errors have been examined primarily to make specific comparisons among models. Having derived these metrics, we place them in context by comparing them to those reported by the reference (TOR) measurement, which include collocated measurement precision and percent of samples below MDL (Table 2).

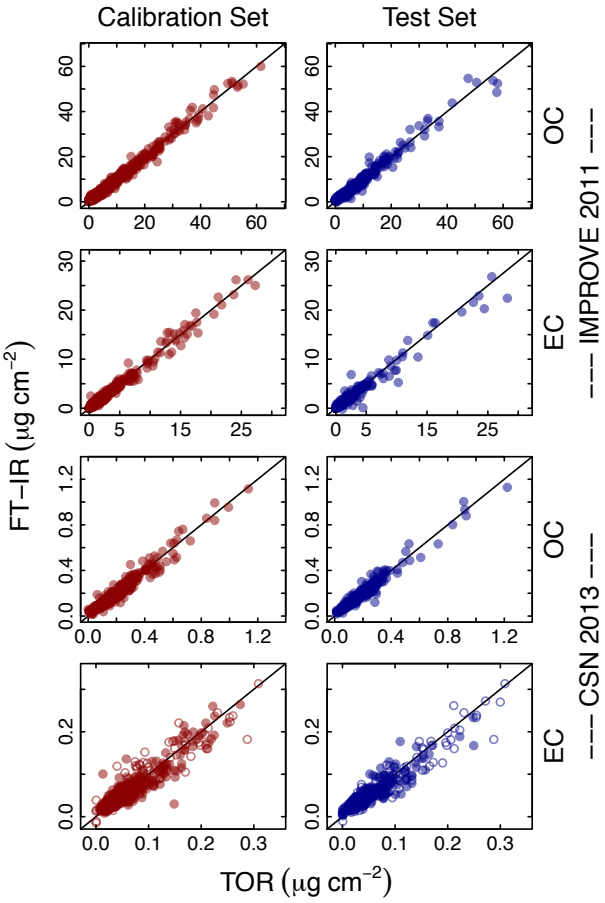

Figure 4. Illustration of model fits ("Calibration Set", left column) and predictions ("Test Set", right column) for the 2011 IMPROVE and 2013 CSN networks. Open circles for CSN EC indicate anomalous samples (discussed in Section 3.5.3). Note units are in areal mass density on the filter.

3.2.2 Systematic errors

In addition to the aggregate metrics discussed above, we evaluate whether essential effects appear to be accounted for in the regression by examining errors across different classes of samples. Systematic patterns or lack of randomness can be evaluated by examining the independence of the individual prediction errors with respect to composition, or using time and location of sample collection as surrogates for composition. For instance, high prediction errors elevated over multiple days may be associated with aerosols of ~~a particular~~unusual composition transported under synoptic scale meteorology that is not well-represented in the calibration samples. Special exception is made for concentration, as errors can be heteroscedastic (i.e., non-constant variance) on account of the wide concentration range of atmospheric concentrations that may be addressed by

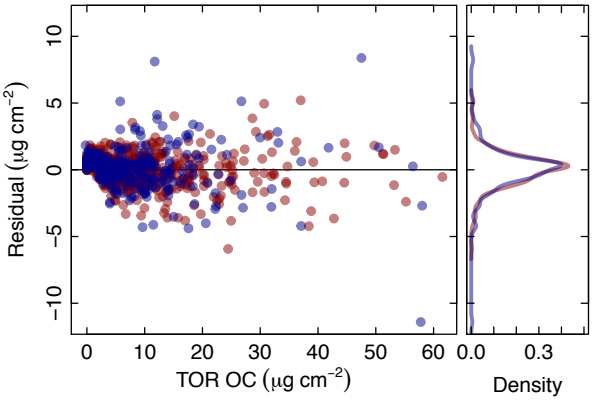

Figure 5. Residuals (red symbols) and prediction errors (blue symbols) from 2011 IMPROVE OC (baseline corrected, base case) predictions. The corresponding kernel density estimate of the distribution is shown on right.

Table 1. Definition for figures of merit for overall assessment of prediction error, samples to which they are applied, and their reference distribution (if available) used for significance testing. $\boldsymbol{y}$ is the (mean-centered) response vector (i.e., TOR OC or EC mass loadings), and $\hat{\boldsymbol{y}}$ is the predicted response (eq. 5). $\langle\cdot\rangle$ is the sample mean, Med$[\cdot]$ is the sample median, and Var$[\cdot]$ is the unbiased sample variance. N_c is the number of paired collocated samples.

Metric	Samples	Estimate	Ref. dist.
root mean square error (RMSE)	all	$\sqrt{\langle(\hat{\boldsymbol{y}}-\boldsymbol{y})^2\rangle}$	χ^2
mean bias	all	$\langle\hat{\boldsymbol{y}}-\boldsymbol{y}\rangle$	
median bias	all	$\mathrm{Med}[\hat{\boldsymbol{y}}-\boldsymbol{y}]$	
mean absolute error	all	$\langle\lvert\hat{\boldsymbol{y}}-\boldsymbol{y}\rvert\rangle$	t
median absolute deviation	all	$\mathrm{Med}[\lvert(\hat{\boldsymbol{y}}-\boldsymbol{y})-\mathrm{Med}[\hat{\boldsymbol{y}}-\boldsymbol{y}]\rvert]$	
coefficient of determination (R^2)	all	$1-(\hat{\boldsymbol{y}}-\boldsymbol{y})^T(\hat{\boldsymbol{y}}-\boldsymbol{y})/\left(\boldsymbol{y}^T\boldsymbol{y}\right)$	F
minimum detection limit (MDL)	blank	$3\sqrt{\mathrm{Var}[(\hat{\boldsymbol{y}}-\boldsymbol{y})]}$	χ^2
collocated precision	collocated	$\lVert\hat{\boldsymbol{y}}_1-\hat{\boldsymbol{y}}_2\rVert/\sqrt{2N_\mathrm{c}}$	t

a single calibration model. This heteroscedasticity leads to a distribution that is leptokurtic (i.e., heavy-tailed) compared to a normal distribution, as shown in Figure 5. As solution algorithms for PLS are agnostic with respect to such residual structure, their application to this type of problem is well-suited.

Given the propensity of prediction error distributions to be long-tailed, error and residual values are transformed to standard-normal variates using inverse hyperbolic sine (IHS) functions (Johnson, 1949; Burbidge et al., 1988; Tsai et al., 2017) using parameters derived from samples with similar analyte (TOR) concentrations. Such a transformation aids identification of systematic errors in prediction related to sample collection time and location; a control chart is displayed for TOR-equivalent OC in Figure 6. Each prediction error is then characterized by its Z-score, which gives an immediate indication of its relation to

Table 2. Description and figures of merit for "base case" models. "Predictors" describe the number of wavenumbers and "Components" describe the number of LVs. Bias and errors are estimated by ensemble medians.

Network	FT-IR	Baseline correction	Wavenumber selection	Predictors	Components
IMPROVE 2011	OC	Spline	None	1563	15
	EC	Raw	None	2784	multilevel
CSN 2013	OC	2nd derivative	BMCUVE	375	3
	EC	Spline	BMCUVE	multilevel	multilevel

Network	FT-IR	R^2	Bias ($\mu g\,m^{-3}$)	Error ($\mu g\,m^{-3}$)	MDL ($\mu g\,m^{-3}$)	Below MDL (%)	Precision ($\mu g\,m^{-3}$)
IMPROVE 2011	OC	0.97	0.01	0.08	0.11	0.7	0.21
	EC	0.96	0.00	0.03	0.01	2	0.06
CSN 2013	OC	0.95	0.04	0.15	0.49	3.0	0.19
	EC	0.88	0.02	0.11	0.17	4.8	0.04

Network	TOR	MDL ($\mu g\,m^{-3}$)	below MDL (%)	Precision ($\mu g\,m^{-3}$)
IMPROVE 2011	OC	0.05	1.5	0.14
	EC	0.01	3	0.11
CSN 2013	OC	0.51	2.7	0.23
	EC	0.03	16.7	0.09

other prediction errors for samples with similar concentrations. Because of the IHS transformation, the magnitude of errors do not scale linearly in vertical distance on the chart, but conveys its centrality, sign, and bounds of the error (e.g., 3 units from the mean encompasses 99% of errors in samples similar in concentration). In this data set, we can see that prediction errors for Sac and Fox (SAFO~~X~~) in each concentration regime are biased positively during the winter, but systematically trend toward the mean toward the summer months. Other high error samples near the 99th percentile (± 3 probits) occur in the urban environment of Phoenix, where the TOR OC concentrations are also highest. However, the prevalence of higher errors in only one of the two Phoenix measurements (PHOE5) may be indicative of sampler differences, rather than unusual atmospheric composition. Errors are negatively biased during the summer months in Trapper Creek, when TOR OC concentrations are typically low.

Systematic errors arising from under-representation of concentration or composition range in the calibration set of IMPROVE was investigated by deliberate permutations of calibration and test set samples by Dillner and Takahama (2015a; 2015b). This study is discussed together with model interpretation (Section 3.5.1). Weakley et al. (2018b) a found systematic errors with respect to OC/EC ratios when predicting TOR-equivalent EC concentrations in the CSN network. These samples were found to originate from Elizabeth, NJ, (ELLA) which differed from the nine other examined sites on account of the high

contributions from diesel PM and extent of reduced charring compared to other samples. The solution was to build a separate calibration model (Section 3.5.3).

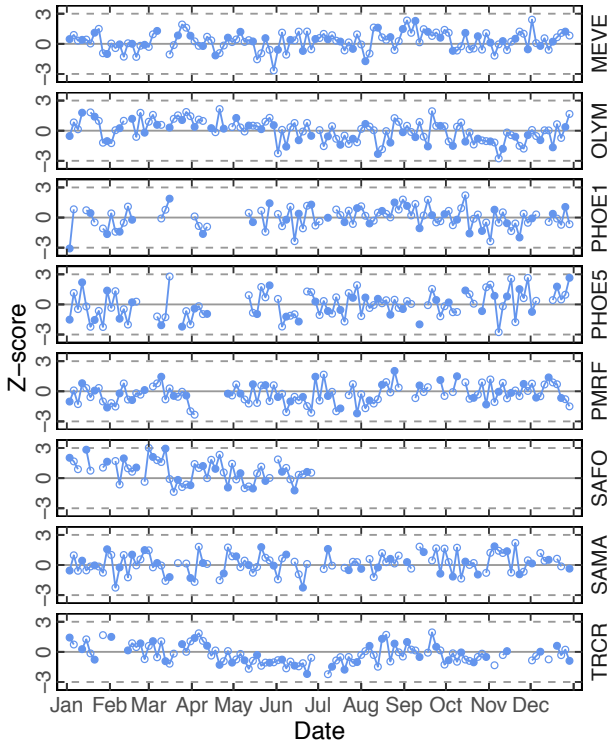

Figure 6. Time series chart of TOR-equivalent OC residuals (for calibration samples) and prediction errors (for test set samples) separated by site. Each value (residual: open circle, prediction error: filled circle) is mapped to a median-centered inverse hyperbolic sine function using 175 values (approximately 20% of the 2011 IMPROVE set) from neighboring TOR OC concentrations to derive distribution parameters so that values are defined within a normal distribution (p-value > 0.2). Dotted horizontal lines indicate ± 3 standard deviations of the standard normal variate (Z-score).

3.3 Spectral preparation

Mid-IR spectra can be processed in many different ways for use in calibration. The primary reasons for spectral processing are to remove influences from scattering such that calibrations models follow the principles of the linear relation outlined in equation 4, and to remove unnecessary wavenumbers or spectral regions that degrade prediction quality or interpretability. Scattering of particles manifests itself in a broad contribution to the signal that is present in the measured spectrum by FT-IR and is addressed by a class of statistical methods referred to as baseline correction (Section 3.3.1). It is even possible to model nonlinear relationships such as the scattering contribution to the signal by a linear model with additional LVs, but these phenomena may not be mixed together with the noise (Borggaard and Thodberg, 1992; Despagne and Luc Massart,

1998). Elimination of unnecessary wavenumbers can reduce noise in the predictions and confer interpretation on the important absorption bands used for prediction; the class of procedures used in this is referred to as variable selection, uninformative variable elimination, among other names (Section 3.3.2). Some algorithms can separate the influence of the background and select variables in the process of finding the optimal set of coefficients $\boldsymbol{b}$ in eq. 5. In each of the following sections, the each of the topics in spectral processing will be introduced before describing their applications to TOR calibrations.

3.3.1 Baseline correction

Baseline correction can be fundamental to the way spectra are analyzed quantitatively. Significant challenges exist in separating the analyte signal from the baseline of mid-IR spectra, which include the superposition of broad analyte absorption bands (O-H stretches in particular) to the broadly varying background contributions from scattering. The algorithm for baseline correction may therefore depend on the type of analyte and the broadness of its profile; optimization of the correction becomes more important as concentrations decrease such that they become difficult to distinguish from the baseline. Approaches can be categorized as reference-dependent or reference-independent (Rinnan et al., 2009), and can be handled within or outside of the regression step. Reference-dependent methods define the baseline with respect to an external measurement, which may be a reference spectrum (Afseth and Kohler, 2012) or concentrations of an analyte. For instance, orthogonal signal correction (OSC) (Wold et al., 1998) isolates contributions to the spectrum that are uncorrelated with the analyte, and can be conceptualized as containing baseline effects. OSC can be incorporated into PLS in which the orthogonal contribution would be represented by underlying LVs (Trygg, 2002). Even without explicit specification of orthogonal components, the influence of baseline effects is accounted for by multiple LVs in the standard PLS model (Dillner and Takahama, 2015a). Reference-independent baseline correction methods remove baseline contributions based on the structure of the signal without invocation of reference values. Two examples described below include interpolation and derivative correction methods. A more comprehensive discussion on this topic is provided by Rinnan et al. (2009).

While theories for absorption peak profiles are abundant, the lack of corollaries for baselines (Dodd and DeNoyer, 2006) lead to semi-empirical approaches for modeling their effects. If we conceptualize the broad baseline as an N-th order polynomial, we can approximate this expression with an analytical function or algorithm. Models can be considered to be (globally) parameteric (e.g., polynomial, exponential) across a defined region of a spectrum, or non-parameteric (e.g., spline or convex hull; Eilers, 2004) in which case local features of the spectrum are considered with more importance. These approaches typically determine the form of the curve by training a model on regions without significant analyte absorption, and interpolated through the analyte region. The modeled baseline is then subtracted from the raw spectrum such that the analyte contribution remains. Model parameters are selected such that processed spectra conform to physical expectations — namely, that blank absorbances are close to zero and analyte absorbances are non-negative. In general, these approaches aim to isolate the absorption contribution to the spectra that are visually recognizable, and therefore most closely conform to traditional approaches for manual baseline removal used by spectroscopists. In addition to quantitative calibration or factor analytic applications (e.g., multivariate curve resolution, de Juan and Tauler, 2006), these spectra are more amenable for spectral matching.

Alternatively, taking the first n-th derivatives of the spectrum will remove the first n terms of the N-th order polynomial and transform the rest of the signal (DeNoyer and Dodd, 2006). Since Gaussian (and most absorption) bands are not well-approximated by low order polynomials, they are not eliminated, i.e., their relative amplitudes and half-widths (ideally) remain unaffected by the transformation. This ensures that their value is retained for multivariate FT-IR calibrations (Weakley et al., 2016). Moreover, derivative-based methods can improve resolution of absorption bands after transformation (illustrated in Figure 7). Derivative transformations can affect the signal-to-noise (S/N) ratio, however; inflating the relative contribution of small perturbations. Therefore, smoothed derivative methods such as the three-parameter Savitzky-Golay filter (Savitzky and Golay, 1964) are favored in order to minimize this effect and, in practice, only first and second derivatives are generally used with vibrational spectra to maintain a reasonable S/N ratio (Rinnan, 2014). In complex aerosol spectra caution must exercised when interpreting the bands resolved by smoothed derivative filters since the filter parameters (i.e., bandwidth, kernel) all influence the outcome of the transformation. A major disadvantage of derivative filtering, in addition to the reduced visual connection to the original spectrum, relates to the inadvertent removal of broad absorption bands (Griffiths, 2006). Tuning filter parameters by trial-and-error may limit this type of band suppression to some extent. As a rule of thumb, the broad O-H stretches of alcohols (3650–3200 cm^{-1}), carboxylic acids (3400–2400 cm^{-1}), and N-H stretches of amines (3500–3100 cm^{-1}) are likely to be sacrificed as a result of derivative filtering (Shurvell, 2006). A willingness to balance this type of information loss against the simplicity and rapidity afforded by derivative methods must be considered in practice.

Different approaches have been used for processing of spectra for TOR calibration, including two interpolation and one derivative approach. Spectral processing is useful for spectra of PM collected on PTFE filters due to the significant contribution of scattering from the PTFE (McClenny et al., 1985). Small differences in filter characteristics lead to high variation in its contribution to each spectrum; a simple blank subtraction of similar blank filters or the same filter prior to PM loading is not adequate to obtain spectra amenable for calibration (Takahama et al., 2013). As the magnitude of this variability is typically greater than the analyte absorbances, baseline correction models trained on a set of blank filters typically do not perform adequately in isolating the non-negative absorption profile of a new spectrum. Accurate predictions made by PLS without explicit baseline correction suggest that the calibration model is able to incorporate its interferences effectively within its feature space if trained on both ambient samples and blank samples together, though visually interpretable spectra for general use is not necessarily retrievable from this model. For this purpose, models based on interpolation from the sample spectrum itself has been preferred. Takahama et al. (2013) described semi-automated polynomial and linear fitting to remove PTFE residuals remaining from blank-subtracted spectra, which was based on prior work for manual baseline correction by Maria et al. (2003) and Gilardoni et al. (2007). This correction method had been used for spectral peak-fitting, cluster analysis, and factor analysis (Russell et al., 2009; Takahama et al., 2011) previously, and was used for 2011 IMPROVE TOR OC and EC calibration shown in Table 2 (Dillner and Takahama, 2015a, b; Takahama et al., 2016). Kuzmiakova et al. (2016) introduced a smoothing spline method which produced similar baseline corrected spectra (both visually and with respect to clustering and calibration) in ambient samples to the polynomial method without need for PTFE blank subtraction. While the non-analyte regions of the spectra are implicitly assumed, the flexibility of the local splines combined with an iterative method for readjusting the non-analyte region effectively reduced the number of tuning parameters from four (in the global polynomial

approach) to one. The spline baseline method was used for TOR EC prediction in 2013 CSN (Weakley et al., 2018b). Second derivative baseline correction method was applied to 2013 CSN TOR OC calibration (Weakley et al., 2016).

Overall, differences in calibration model performance in TOR prediction between spline corrected and raw spectra models were minor for the samples evaluated in 2011 IMPROVE (results were comparable to metrics in Table 2). However, wavenumbers remaining after uninformative ones were eliminated (Section 3.3.2) differed when using baseline corrected and raw spectra — even while the two maintained similar prediction performance. Weakley et al. (2016) and Weakley et al. (2018b) used the Savitzky-Golay method and spline correction method for TOR OC and EC, respectively, in the 2013 CSN network, but did not systematically investigate the isolated effect of baseline correction on predictions without additional processing. A formal comparison between the derivative method against raw and spline-corrected spectra has not been performed, but this is an area warranting further investigation. Standardizing a protocol for spectra correction based on targeted analyte is a sensible strategy, as spectral derivatives are associated with enhancement in specific regions of the spectra. The selection of baseline correction method may also consider the areal density of the sample, since the S/N is reduced with derivative methods. However, the success of derivative methods demonstrated for TOR OC in CSN samples (with systematically lower areal loadings than IMPROVE samples) indicates that the reduction in S/N is not likely a limiting factor for quantification in this application.

The derivative method appears to have significant advantage in reducing the number of LVs as demonstrated for TOR OC (Table 2). The derivative-corrected spectra model for 2013 CSN resulted in only 4 components in contrast to the 35 selected by the raw spectra model. While wavenumber selection and a different model selection criterion was simultaneously applied to the derivative-corrected model, a large reason for the simplification is likely due to the baseline correction. For reference, reduced-wavenumber raw spectra models for 2011 IMPROVE TOR OC and EC still required 7–9 components (the full-wavenumber model required 15–28, depending on spectral baseline correction) (Takahama et al., 2016). A parsimonious model is desirable in that it facilitates physical interpretation of individual LVs as further discussed in Section 3.4.

The effect of baseline correction on reducing the scattering is illustrated by revisiting the TOR-equivalent OC predictions for the 2013 IMPROVE data set. Reggente et al. (2016) found that the raw spectra 2011 IMPROVE calibration model performed poorly in extrapolation to two new sites in 2013, particularly FRES and BYIS. When using baseline corrected spectra, the median bias and errors are reduced from 0.28 $\mu g\,m^{-3}$ and 0.43 $\mu g\,m^{-3}$ and to 0.19 $\mu g\,m^{-3}$ and 0.28 $\mu g\,m^{-3}$, and R^2 increases from 0.79 to 0.91 for samples from these sites (Figure for baseline corrected predictions shown in Section 4.1.1). As the filter type remained the same, this improvement in prediction accuracy is likely due to the removal of scattering contributions in $PM_{2.5}$ particles in the new set that differs from the calibration set. Spectral signatures of nitrate and dust suggested the presence of coarse particles different than those in the 2011 calibration (and test) set samples. 4.1).

3.3.2 Wavenumber selection

Wavenumber or variable selection techniques aim to improve PLS calibrations by identifying and using only germane predictor variables (Balabin and Smirnov, 2011; Höskuldsson, 2001; Mehmood et al., 2012). Typically, such techniques remove variables deemed excessively redundant, enhance the precision of PLS calibration, reduce collinearity in the variables (and therefore model complexity) (Krämer and Sugiyama, 2011), and possibly improve interpretability of the regression. The sim-

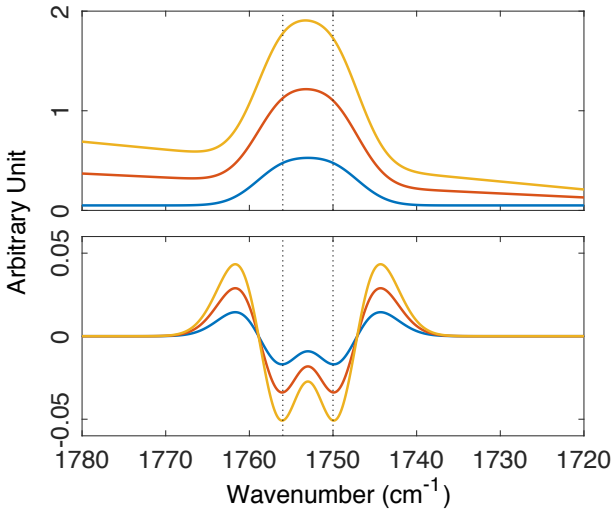

Figure 7. Three synthetic absorption spectra constructed with varying contributions from a polynomial baseline and two unresolved Gaussian peaks (A), and their 2nd order, 5-point, second derivative, Savitzsky-Golay filter transformations (B). Absorption spectra were constructed such that the additive, linear, and polynomial components of the baseline scale with the amplitude of the absorption bands. ~~The baseline is completely eliminated and the bands are better resolved as a result of filtering.~~

plest variable selection method based on physical insight rather than algorithmic reduction is truncation, in which regions for which absorbances are not expected or expected to be uninformative are removed a priori. Algorithmic variable selection techniques fall into three categories: filter, wrapper, and embedded methods (Saeys et al., 2007; Mehmood et al., 2012).

Filter methods provide a one-time (single-pass) measure of a variable importance with important and redundant variables distinguished according to a reliability threshold. Variables above such a threshold are retained and used for PLS calibration. Often, thresholds are either arbitrary or heuristically determined (Chong and Jun, 2005; Gosselin et al., 2010). In general, filter methods are limited by their need to choose an appropriate threshold prior to calibration, potentially leading to a suboptimal subset of variables.

The essential principle of wrapper methods is to apply variable filters successively or iteratively to sample data until only a desirable subset of quintessential variables remain for PLS modeling (Leardi, 2000; Leardi and Nørgaard, 2004; Weakley et al., 2014). Wrappers operate under the implicit assumption that single-pass filters are inadequate, requiring a guided approach to comprehensively search for the optimal subset of modeling variables. Since searching all $2^p - 1$ combinations of wavenumbers is not tractable for multivariate FT-IR calibration problems ($p > 10^3$), model inputs (or importance weights) are generally randomized at each pass of the algorithm to develop importance criteria, foregoing an exhaustive variable search. Genetic algorithms and backward Monte Carlo unimportant variable elimination (BMCUVE) are examples of two randomized wrapper methods (Leardi, 2000; Leardi and Nørgaard, 2004). Wrapper methods generally perform better than simple filter methods and have an additional benefit of considering both variables and PLS components simultaneously during optimization. The major

drawback to wrapper methods are generally longer runtimes (which may be on the order of hours for large-scale problems) than filter methods.

As their name implies, embedded methods nest variable selection directly into the main body of the regression algorithm. For example, sparse PLS methods (SPLS) eliminate variables from the PLS loading weights ($\boldsymbol{w}$), which reduce the number of non-zero regression coefficients ($\boldsymbol{b}$) when reconstructed through eq. 5 (Filzmoser et al., 2012). The zero-valued coefficients obtained for each LV can possibly confer component-specific interpretation of important wavenumbers, but leads to a set of regression coefficients which are overall not as sparse as methods imposing sparsity directly on the regression coefficients (Takahama et al., 2016).

Many methods select informative variables individually, but for spectroscopic applications it is often desirable to select a group of variables associated with the same absorption band. Elastic net (EN) regularization (Friedman et al., 2010) adds an L2 penalty to the regression coefficient vector in addition to the L1 penalty imposed by the least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996), thereby imparting a grouping effect in selection. Interval variable selection methods (Wang et al., 2017) draw upon methods discussed previously but employ additional constraints or windowing methods to target selection of contiguous variables (i.e., an algorithmic approach to truncation).

Takahama et al. (2016) evaluated two embedded (sparse PLS) algorithms and one hyphenated method in which EN used as a filtering method prior to PLS calibration (EN-PLS, Fu et al., 2011) for TOR OC and EC calibration in the IMPROVE network. A suite of reduced-wavenumber models were considered by varying model parameters that controlled the sparsity, and evaluated using cross-validation and separate test set samples. Since full-wavenumber calibration models (both raw and baseline corrected) for TOR OC and EC in the IMPROVE networks already performed well (Section 3.2.1), wavenumber selection did not improve model predictions but served mostly to aid interpretation of the most important absorption bands. Takahama et al. (2016) found that these methods could use as little as 4–9% of the original wavenumbers (2784 for raw and 1563 for spline corrected) to predict TOR-equivalent OC and EC. EN-PLS consistently achieved the sparsest solution (by more than a factor of two in almost all cases) on account of the LASSO penalty applied directly to the regression vector. While all variable selection methods generally performed well for TOR-equivalent OC and EC prediction in 2011 IMPROVE samples, calibrations for organic functional groups built using sparse PLS algorithms appeared to be less robust in extrapolation to ambient sample spectra. While also being the most sparse, EN-PLS yielded similar predictions to the original PLS (full wavenumber) models (Takahama and Dillner, 2015) that led to OC reconstruction from summed functional group contributions having better agreement with TOR OC than other sparse calibration algorithms, including EN without PLS. This finding suggests that variables eliminated for being uninformative in the calibration set samples may lead to undesirable oversimplification of a model that may be used with samples with potentially different composition, though this hypothesis has yet to be tested with calibrations developed with ambient measurements as reference, where the extent of extrapolation may not be so severe as with calibrations developed with laboratory standards. Weakley et al. (2016, 2018b) applied BMCUVE to second derivative or spline corrected spectra in the CSN network. Improved MDL but otherwise similar performance metrics to the raw (full wavenumber) calibration model was obtained using the reduced model for TOR OC (performance described in Section 3.2.1), though the individual contributions of baseline correction and wavenumber selection to improvement in MDL was not investigated. The impact of

wavenumber selection on model performance was not investigated for TOR EC, but the reduced-wavenumber model predicted EC within TOR precision (Section 3.2.1). Interpretation of the selected wavenumbers are discussed in Section 3.4.

3.4 Interpretation of important variables and their interrelationships

Interpreting the relationships among variables being used by a statistical model to make predictions is a challenging topic on account of their semi-empirical basis. In particular, it is possible to exploit statistical correlations among the variables to make predictions, which can be detrimental if the correlation changes or model is applied in a different context ~~(further discussion is provided in Section 4.1.2)~~. Therefore, model interpretation is strongly related to anticipation of model applicability and a priori identification of samples with potentially high prediction errors (Section 4.1.2). Inspection of how LVs and absorption bands are used by a model can give an indication of their importance, and possibly establish a physical basis between analyte concentrations and their relevant vibrational modes. Existence of sample subgroups and potentially influential subgroups can initiate indentification of relevant sample characteristics that have a disproportionate role in prediction. To some extent, discussions in Sections 3.1 and 3.3.2 focusing on eliminating uninformative variables (LVs or wavenumbers) during the model selection process is also relevant in this context (some of the same techniques are applicable to both tasks), but the focus will be on understanding the importance of the remaining variables. The importance of samples and specific attributes (concentration or composition) associated with them are addressed in Section 3.5.

As with complex mechanistic models, a general investigation can be carried out through sensitivity analyses (Harrington et al., 2000; Chen and Yang, 2011). One of the advantages of a PLS regression approach is that the contribution of each LV to response ($\boldsymbol{y}$) or spectra matrix ($\boldsymbol{X}$) can be characterized by the explained Sum-of-Squares (*SS*) and its normalized surrogate, Explained Variation (*EV*) (Martens and Næs, 1991; Abdi, 2010). The emphasis placed by a model on particular wavenumbers can be examined through its regression coefficients $\boldsymbol{b}$, Selectivity Ratio (*SR*) (Kvalheim, 2010), or the Variable Importance in Projection (*VIP*) metric (Wold, 1993; Chong and Jun, 2005). These quantities can be written using j and k as indices for wavenumber and LV (with J as the total number of wavenumbers), respectively:

$$SS_{y,k} = q_k^2 \boldsymbol{t}_k^T \boldsymbol{t}_k \tag{8}$$

$$SS_{X,k} = (\boldsymbol{p}_k^T \boldsymbol{p}_k) \cdot (\boldsymbol{t}_k^T \boldsymbol{t}_k) \tag{9}$$

$$SS_{X,j} = \boldsymbol{p}_j \left(\boldsymbol{T}^T \boldsymbol{T} \right) \boldsymbol{p}_j^T \tag{10}$$

$$EV_{y,k} = SS_{y,k} / (\boldsymbol{y}^T \boldsymbol{y}) \times 100\% \tag{11}$$

$$EV_{X,k} = SS_{X,k} / (\boldsymbol{X}^T \boldsymbol{X}) \times 100\% \tag{12}$$

$$SR_j = SS_{X,j} / \left(\boldsymbol{e}_{X,j}^T \boldsymbol{e}_{X,j} \right) \tag{13}$$

$$VIP_{jk} = \left(J \frac{\sum_{\ell=1}^{k} SS_{y,\ell} \left(w_{\ell j} / \|\boldsymbol{w}_\ell\| \right)^2}{\sum_{\ell=1}^{k} SS_{y,\ell}} \right)^{1/2} . \tag{14}$$

~~These expressions apply to both calibration and new samples, though typically used for the former.~~ Note that for new samples, the loadings ($\boldsymbol{q}$ and $\boldsymbol{p}$), sum-of-squares ($\boldsymbol{T}^T \boldsymbol{T}$) and the means used for centering of each array ($\boldsymbol{y}$ and $\boldsymbol{X}$) are fixed according

to the calibration set. For PLS, the EV_X is not as commonly examined as for other factor analysis techniques as the primary objective is in explaining the variation in $\boldsymbol{y}$. In addition to metrics characterizing the overall importance of latent and physical variables, the (normalized Euclidean) distance of individual samples from the center of the calibration space can be indicated by its leverage h. For mean-centered PLS, h is computed for row vector of new scores $\boldsymbol{t}$ corresponding to sample i weighted by the inverse sum-of-squares of the calibration set (Martens and Næs, 1991):

$$h_i = \boldsymbol{t}_i \left(\boldsymbol{T}^T \boldsymbol{T} \right)^{-1} \boldsymbol{t}_i^T \tag{15}$$

The sample leverage is used to assess influential points in the model, identify outliers, and estimate prediction variance (prediction intervals). Further discussion of leverage used in the last two objectives is discussed in Section 4.1. Regression coefficients can oscillate between positive and negative numbers as higher number of LVs are used (Gowen et al., 2011) and their magnitude must be considered together with that of the absorbance (i.e., large regression coefficients coupled with small absorbances may not have a large impact on the modeled outcome), metrics such as SR or VIP can be more useful to assess their relative importance (the two vary in ease of interpretability for different types of data and data sets, Farrés et al., 2015).

For TOR analysis, VIP scores have been used to interpret wavenumber importance (Dillner and Takahama, 2015a, b; Weakley et al., 2016, 2018b). VIP scores can also be used as a filtering method (Section 3.3.2) for wavenumber selection (e.g., Gosselin et al., 2010; Lin et al., 2013; Liu, 2014), but here they have been used only for post hoc interpretation for this work. The main principle is that the mean VIP score across all wavenumbers is unity, so those with more influence in explaining $\boldsymbol{y}$ carry values above and those with less influence fall below. However, Chong and Jun (2005) found that the actual importance threshold can be data-specific, with dependence on the proportion of uninformative predictors, predictor correlation, and the actual values of the regression coefficients. Meaningful threshold values varied between 0.8 and 1.2 in their work ~~of Chong and Jun (2005)~~. VIP scores for TOR models are summarized in Figure 8. Wavenumbers associated with TOR OC not surprisingly span~~s~~ a range of functional group structures. Common functional groups interpreted for both 2011 IMPROVE and 2013 CSN include aliphatic C-H and carbonyls (carboxyl, ketone, ester, aldehyde), with possible contributions from various nitrogenated (amine, amide, nitro) groups (Takahama et al., 2016; Weakley et al., 2016). Other candidate bonds are described but assigned with less certainly on account of strong overlap of absorption bands in some spectral regions. Takahama et al. (2016) based their interpretation on the selected wavenumbers and VIP scores for both raw and baseline corrected models under a "common bond" that the two models are basing their prediction using the same set of functional groups rather than different ones. Based on this assumption, it appeared that the two models were using different vibrational modes (stretching or bending) for aliphatic C-H and alcohol O-H, though bending modes typically exhibit weaker absorption signatures. The capability to accurately predict TOR-equivalent OC concentrations in samples with different OM/OC ratios (determined by functional group calibration models with FT-IR) as discovered through permutation analysis (Section 3.5.1) suggests that on average, there is some insensitivity to weighting of functional groups that determine the degree of functionalization in the sample.

For TOR EC, among other functional groups, wavenumbers selected between 1600–1500 cm^{-1} were attributed to C-C and C=C stretching in skeletal ring structures of aromatic or graphitic carbon (Takahama et al., 2016; Weakley et al., 2018b). While this absorption band corresponds to lattice vibrations in graphitic carbon (Tuinstra and Koenig, 1970) and commonly used

in Raman spectroscopy for characterization of soot particles (Sadezky et al., 2005; Doughty and Hill, 2017), a peak has been observed in mid-IR spectra only after crystalline structure is broken down through mechanical stress (Friedel and Carlson, 1971, 1972; Ţucureanu et al., 2016). Nonetheless, a peak of moderate to broad width in this region is observed in soot (Akhter et al., 1985; Kirchner et al., 2000; Cain et al., 2010), soil black carbon (Bornemann et al., 2008; Cheng et al., 2008), and coal (Painter et al., 1982). In constructing a PLS model to predict BC in soil by mid-IR spectra and PLS, Bornemann et al. (2008) further removed the potential influence of correlation between EC and OC in soil samples by predicting the BC content normalized by OC with an R^2 of 0.81. This analysis encouraged their interpretation that the aromatic structures visible in their first PLS loading weight vector were specific to BC, which potentially supports the same interpretation for atmospheric samples. However, Weakley et al. (2018b) found that a calibration model for ELLA did not require aromatic structures for prediction of TOR-equivalent EC. This site was located in close proximity to a toll station on the New Jersey turnpike and was characterized by high diesel PM loading, low OC/EC ratio, and low degree of charring compared to samples from other CSN sites in the 2013 data set. The calibration model was able to predict TOR-equivalent EC concentrations primarily using absorption bands associated with aliphatic C-H (also selected in the calibration model for the other 2013 CSN sites) and nitrogenated groups believed to be markers for diesel PM. A standard method for quantification of soot (ASTM D7844-12, 2017) recommends the use of scattering characterized at 2000 cm^{-1} (without baseline correction) on the assumption that there is no absorption usable for quantification. Given that baseline corrected spectra (in which scattering at 2200-1900 cm^{-1} in addition to other wavenumbers with negligible absorption are forced to zero) are able to predict TOR-equivalent EC concentrations in both 2011 IMPROVE and 2013 CSN — and most relevant wavenumbers are in regions associated with visible absorption peaks — the predictions do not appear to be based on scattering in this application. Early work by Pollard et al. (1990) reported a calibration for collocated EGA EC using a peak located at 666–650 cm^{-1} in mid-IR spectra of PM collected onto PTFE filters at Glendora, CA. However, what vibrational mode this peak corresponds to is unclear, as there is also IR interference from the PTFE substrate in this region (Quarti et al., 2013). The true nature of operationally-defined TOR EC and a definitive reason that its concentration can be predicted from mid-IR spectra is an ongoing topic of investigation. Surface functionalization of graphitic combustion particle surfaces (Cain et al., 2010; Popovicheva et al., 2014) are estimated to be a small fraction of the functional groups from organic aerosol in the same sample, and therefore considered to be unlikely to be useful for calibration. Soot emissions comprise both light-absorbing black carbon and organic carbon (Novakov, 1984; Petzold et al., 2013), and it is possible that both fractions exhibit mid-IR activity (some structures co-absorbing in the same region) that can be used for quantification. Whether the functional groups used for prediction of TOR-equivalent EC are due to the organic fraction associated with incomplete combustion or other indirect markers warrants further investigation in controlled studies.

While the large number of LVs used by the IMPROVE calibration models precluded attempts at identification of individual components, Weakley et al. (2016) was able to do this for 2013 CSN TOR OC calibration models on account of their low complexity. Application of second-derivative baseline correction, BMCUVE wavenumber selection, and model selection by Wold's R criterion resulted in a 4-LV model for TOR OC. Further nuanced interpretation was aided by re-projection of LVs onto PCA space which modeled much of the same variance as PLS scores, but were formulated and arranged according to their capability to explain the remaining variance in the spectra instead of the covariance with respect to TOR OC. By visualizing

the sample spectra in two dimensions of this space using a conventional biplot, Weakley et al. (2016) identified a subset of samples with extraneous variance in 2013 CSN spectra attributed to water vapor in the beam path present during spectra acquisition in the laboratory. While the water vapor conferred minimal prediction error, loading this spectral interference onto one dimension and excluding it the final calibration model improved interpretability with a more parsimonious model using only the 3 remaining components. Surprisingly, a single component representing an organic mixture explained close to 90% of the TOR OC variance, with the remaining two components attributed to interferents: PTFE substrate and ammonium nitrate (explained variation of 3–4 % each).

Model interpretation is a continual challenge but a necessary aspect of statistical modeling from a chemometrics perspective, and remains an active area of investigation for TOR analysis. While the LVs are not constrained to be non-negative as factors for Multivariate Curve Resolution, Positive Matrix Factorization and Non-negative Matrix Factorization (Paatero, 1997; Lee et al., 1999; de Juan and Tauler, 2006), the relative variation of scores can be analyzed alongside auxiliary measurements to identify their importance toward specific PM samples. This association can be made in a correlative capacity (Russell et al., 2009; Faber et al., 2017), or through more sophisticated means such as target transformation factor analysis (Henry et al., 1984; Hopke, 1989). In addition, the way of obtaining LVs can be modified to accommodate features from TOR OC and EC simultaneously. A variant of PLS that can potentially aid in this endeavor is "PLS2", which uses a shared representation of LVs for multiple response variables (Martens and Næs, 1991). Shared representations are commonly used in multi-task learning (Caruana, 1997) to build models that generalize from fewer, diverse training instances, and may additionally confer benefit in this context for understanding the inter-relationship between these two substances and their thermal fractions. The univariate-response formulation of PLS ("PLS1") as described in Section 3.1 has been the focus of past work with TOR calibrations as it typically achieves the same or better accuracy as PLS2 with fewer LVs (Martens and Næs, 1991), but the potential for PLS2 in improved interpretation and robustness in a wider range of contexts is an area that can be further explored.

3.5 Sample selection

To design a campaign to collect both FT-IR spectra and reference measurements or to select among available collocated measurements in a database to construct a new calibration model, it is necessary to address the question of *how many of which type of samples do we need*? Provided that the form of a data set can be fit by several models, it is possible for the simpler ones with more training data to outperform more complex ones with less training data for new predictions (Halevy et al., 2009). This argument can be rationalized in a chemometric context by conceptualizing an ideal calibration model as one built upon samples of identical composition and concentration (with replicates) for every sample in the prediction set. Especially for complex PM components such as TOR OC and EC that have a multitude of absorption bands in the IR from both target and interfering substances, enough samples must be included in the calibration set to span the range of multiple attributes. For each unique sample removed from the calibration set, the corresponding composition in the prediction set must be estimated by mathematical interpolation or extrapolation from the remaining samples. Reducing the number of calibration samples increases the dependence of the predictions on the functional form or weighting scheme (with respect to variables and samples) of the selected model with possible consequences for prediction accuracy. Lacking mechanistic constraints,

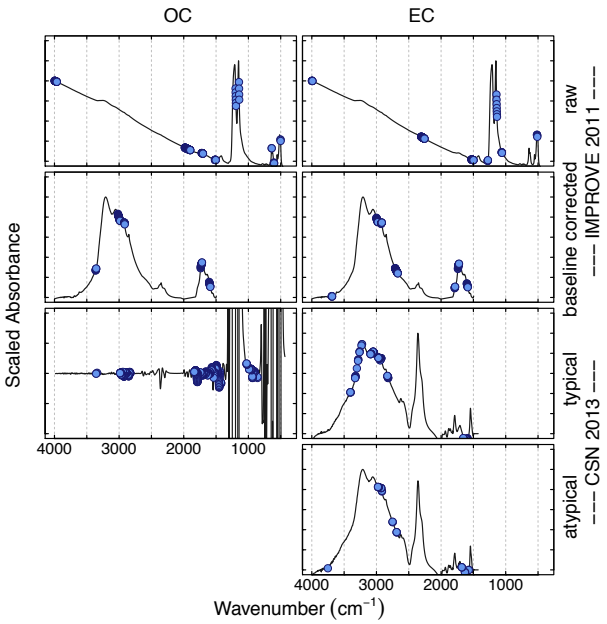

Figure 8. Selected wavenumbers (blue points) overlayed on mean of calibration spectra (black lines). 2011 IMPROVE spectra remain unprocessed ("raw") or baseline corrected using smoothing splines ("baseline corrected"), while the 2013 CSN spectra are baseline corrected using the Savitsky-Golay 2nd derivative approach. "Atypical" and "typical" categories for 2013 CSN EC refer to samples for Elizabeth, NJ, and the remaining nine sites, respectively. [Figure has been changed as spline baselines were used instead of 2nd derivatives for CSN EC.]

predictions from data-driven models may exceed physical limits with increasing reliance on the underlying algorithm over measurements. The obvious importance of chemical similarity in calibration can be related back to physical principles that give rise to the observed mid-IR spectrum. First, for any given wavenumber, the absorption scales with analyte abundance — simpler calibration models in analytical chemistry built on this principle dictate that the concentration range covered by calibration samples should bound the concentrations in the new samples so that values are interpolated rather than extrapolated to minimize prediction error. Second, complex absorption profiles arise from heterogeneous broadening of absorption bands in the condensed phase. Therefore, samples with a similar chemical composition to new samples are likely to have similar patterns of absorbance and interferences that can be accounted for by the calibration model.

A basic premise follows that calibration models built with samples having similar spectroscopic profiles, specifically near the most relevant absorption bands, are likely to yield better prediction results for new samples. For analysis of simple mixtures, one common strategy pursued in experimental design is to prepare samples that populate the chemical coordinates (e.g., molar concentrations of its constituent species) of anticipated composition according to Euclidean distance (Kennard and Stone, 1969). However, this procedure does not guarantee that the training and prediction data will have similar distributions in the feature space of an effective calibration model (i.e., similarity may not be best characterized by Euclidean distances). This task is further complicated by the fact that chemical similarity is not easy to define for composite substances (TOR OC) or

chemically ambiguous quantities (TOR EC). Moreover, the samples for calibration at the level of chemical complexity of atmospheric mixtures are typically limited by the availability of collocated measurements (e.g., TOR reference measurements together with sample spectra from PTFE filters).

In the context of these challenges, the canonical ("base case") strategy for TOR OC and EC calibration has been to use space and time as a proxy for composition. A stratified selection approach — in which selected samples are evenly spaced out over a full year at each measurement site — is used construct the calibration set, as there is reasonable expectation that an adequate representation of emission sources and extent of atmospheric processing can be captured. Blank PTFE filter spectra are added to the calibration set and their corresponding reference concentrations are set to zero, as this value is equally valid to the TOR-determined concentration for below-MDL samples. Excluding irregular events (e.g., wildfires), this approach can be effective in building a general calibration model for atmospheric samples and has demonstrated good performance (Section 3.2). However, samples from the same site and season are not strictly required for successful prediction of each new sample. Reggente et al. (2016) demonstrate accurate prediction for a full year of TOR OC and EC concentrations at sites not included in the calibration (also revisited in Section 4.1). The extent to which site, season, local emission, or meteorological regime of a new sample affects prediction depends on how these factors contribute to deviation in chemical composition from calibration samples. We further summarize our efforts in understanding which types of samples are important (Section 3.5.1) and how many samples are needed (Section 3.5.2) for calibration. Lastly, we describe how specialized calibration models can better serve a specific set of samples that are not well-represented in the feature space of all calibration samples (Section 3.5.3).

3.5.1 Important attributes

Our findings indicate that many, though not all, methods for sample selection can lead to an acceptable calibration model as determined by evaluation criteria described in Section 3.2. To investigate which aspects of similarity are important in this regard, Dillner and Takahama (2015a; 2015b) performed permutation analyses on the available set of samples to study how differences between calibration and test set samples influenced prediction errors. Samples were grouped according to values of descriptors chosen to capture the effect of analyte concentration (TOR OC, EC), source and degree of functionalization (OC/EC and OM/OC), and inorganic interferences (ammonium/OC, ammonium/EC). Predictions were evaluated when the distribution of these descriptors represented in the calibration set were selected to be either similar or different to those in the test set. To construct calibration and test sets according to these specifications, samples were arranged in order of a particular attribute. For similar calibration and test set distributions, every third was reserved for the test set while the remainder was used for calibration. To examine the effect of extrapolation with respect to any attribute, the calibration set was constructed from samples with the lowest two-thirds or highest two-thirds of attribute values, and the remainder used for the test set. To examine the effect of interpolation, the highest third and lowest third were used for calibration and predictions made on the middle third of samples. Inadequate representation of any of these variables in the calibration set led to increased errors in model predictions, but with typically low bias in interpolation. TOR OC could be predicted with only marginal increase in bias (median absolute bias of 0.1 $\mu g\,m^{-3}$) and no increase in normalized error ($\sim$10%) even when extrapolating predictions on average three times higher, indicating a calibration that was effectively linear over the range tested (0–8 $\mu g\,m^{-3}$). For samples

varying in OM/OC ratio between 1.4–2.5, normalized error in predicted TOR OC increased from ∼10% when the calibration and test sets were similar to 14–17% when they were forced to diverge according to the segmentation described above, but the predictions remained unbiased. The largest increase in prediction error came when using calibration samples with low ammonium interference (low ammonium/OC ratio) to high ammonium content, with an increase in normalized error of from ∼10% to 24%. For TOR EC, almost every extrapolation scenario resulted in an increase in either bias or normalized error (by 10 to 60 percentage points), suggesting its sensitivity to a large number of sample attributes.

Such permutation analyses permit independent evaluation of attribute importance only to the extent that they are not correlated in the samples. For instance, for 2011 IMPROVE, much of the variability across the entire data set was driven by the two collocated urban sites in Phoenix, AZ, which contained higher concentrations of less functionalized PM in general than the remaining rural sites. However, normalization strategies — e.g., of ammonium by OC or EC — reduced confounding effects. Dillner and Takahama (2015a; 2015b) only tested each univariate case in turn, but multidimensional permutation analysis in which samples are partitioned according to differences across multiple variables for model building and testing may be possible with a large number of samples. Computational resources permitting, bootstrap sampling combined with post analysis may provide another means of testing the importance of particular attributes in such instances.

3.5.2 Number of samples

The minimum number of samples required by a model is dependent on the capacity of its calibration samples to collectively represent the diversity of composition in new samples, and the algorithm to effectively interpolate or extrapolate into unpopulated regions of the composition space. To illustrate this notion, we present the change in prediction metrics for TOR-equivalent OC as a function of the number of ambient samples in the calibration set (Figure 9). Beginning with samples selected according to the base case strategy (stratifying by space and time) as the initial reference, the number of ambient samples in the calibration set are reduced while the number of blank samples are held constant. The set of test samples are also fixed for all evaluations. While the conclusions are not strikingly obvious, some overall trends can be noted. Figure 9 shows a general decrease in prediction accuracy with fewer number of ambient samples, especially below ∼150 samples, though individual differences among most models are not statistically significant. The gradual degradation in prediction accuracy is attributed to difficulty in maintaining representativeness of important attributes with a small number of samples. Figure 10 shows the increasing difference in empirical probability distributions of attributes in the calibration and test set samples as a function of the number of ambient samples using the Kolmogorov-Smirnov test statistic (higher values indicate higher dissimilarity between the calibration and test set distributions). The increase in differences between the distributions in TOR OC, but particularly the ammonium/OC ratio, is the primary cause as it was determined to be a critical attribute for TOR OC prediction (Section 3.5.1). Due to the diminising statistical power with fewer calibration samples, statistical significance is not established in this regime; we therefore interpret these results qualitatively. The MDL is generally maintained or improved with decreasing number of ambient samples, which is sensible as the number of blank samples grows in proportion. On the other hand, the number of blank samples (varied between 0 and 36) when included with 501 ambient samples in the calibration set (Dillner and Takahama, 2015a, b) did not have a large effect on the MDL.

We might conclude that larger calibration sets that more likely cover the range of attributes in new samples might lead to better model performance. Reggente et al. (2016) shows an example for raw spectra. Without baseline correction, TOR OC concentrations for two sites — FRES and BYIS —in 2013 IMPROVE were not predicted well by the original model. Predictions were shown to improve when samples from these sites were included (Reggente et al., 2016). In this case, the calibration set without ~~Fresno~~FRES and ~~Korea~~BYIS was too small in that it did not contain the appropriate representation of specific sample characteristics.However, as with wavenumbers, populating the calibration set with increasing number of unrelated or uninformative samples with respect to a targeted class of samples may lead to added noise or bias from unfavorable model weighting. In such instances, smaller, dedicated models may be better for specific classes of samples provided that it is possible to distinguish which model is best suited for each sample. In the next section, we describe cases in which a smaller subset of samples for calibration have been found to be appropriate for improving specific performance targets.

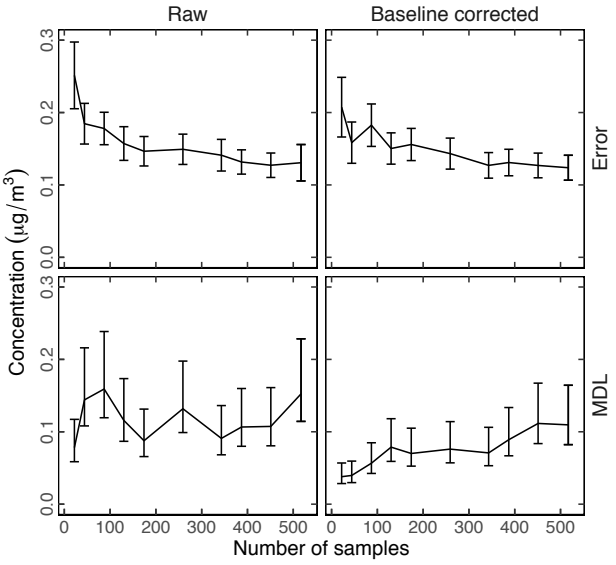

Figure 9. The prediction accuracy for TOR OC as a function of the number of ambient samples in the calibration set (the number of blanks were kept constant at 36). Using the 2011 IMPROVE base case calibration model, every nth sample was removed (which leaves spatial and temporal representation of samples close to the original set). The performance metrics are computed on the same 286 test set samples for all calibration models.

3.5.3 Smaller, specialized models

While a large, monolithic model may be most capable of accommodating diverse composition in prediction set samples, models that assume underlying structure of the chemical domain for interpolation or extrapolation may be susceptible to undue influence by one or more groups of (high leverage) samples and return biased predictions for a specific set of underrepresented samples. Statistical localization is the process by which calibration models are built with samples that are closest in composition

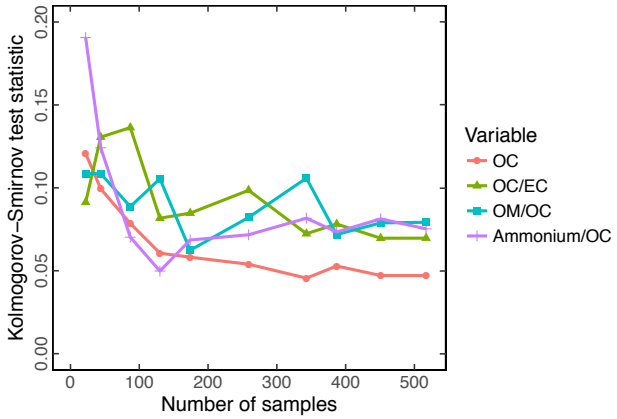

Figure 10. Kolmogorov-Smirnov (KS) test statistic for different number of calibration samples used in Figure 9. The KS statistic characterizes the difference between two empirical distribution functions; in this case determined for probability distributions of each variable between the calibration and test set samples.

to samples for which predictions are desired. While the overall number of samples used for training in each localized model is reduced, the distribution of the calibration model better reflects that of the subset of samples for which new predictions are to be made. Together with a classifier capable of selecting the appropriate localized model for each new spectrum, several models can collectively function as a single multilevel model to provide a best estimate of the targeted concentration.

This approach has been applied to TOR EC calibration in both networks studied (Dillner and Takahama, 2015b; Weakley et al., 2018b) (Figure 11). Dillner and Takahama (2015b) constructed a multilevel model consisting of calibrations for two different concentration regimes for 2011 IMPROVE. A calibration model using only a third of the lowest concentration samples (areal density $<$0.68 $\mu g\,cm^{-2}$) led to an MDL of 0.01–0.02 $\mu g\,m^{-3}$, while using the full range of areal loadings for calibration led to an MDL of 0.03–0.08 $\mu g\,m^{-3}$. Overall prediction errors for low samples were also reduced with a dedicated model, but to a lesser extent than the MDL. The full range model served as a classifier; predictions that fell below the areal loading threshold according to this model were refined with the low-concentration calibration model. As discussed in Section 3.4, ELLA was believed to be influenced by diesel emission sources that led to different PM composition and spectral characteristics from the remaining nine CSN sites. Therefore, predicted concentrations for ELLA were systematically biased low compared to observations. Weakley et al. (2018b) trained a partial least squares discriminant analysis (PLS-DA) model on geographical location to segregate typical samples from atypical ones that resembled ELLA spectra. Spectra classified as being atypical were predicted using a model trained solely on ELLA samples, while the ones classified as typical were predicted using a model trained on the rest of the samples. Considering the overall model performance for all samples, using this multilevel approach led to an improvement in R^2 from 0.76 to 0.88, and a decrease in bias from 5.2 to 2.7% (with corresponding improvements in MDL, precision, and other figures of merit). The difference in metrics were largely due to improvement in ELLA predictions, as the predictions for non-ELLA samples were similar in both approaches (mean errors of 0.15 and 0.16 $\mu g\,m^{-3}$, and R^2 of 0.83 and 0.85 for the monolithic and multilevel model, respectively).

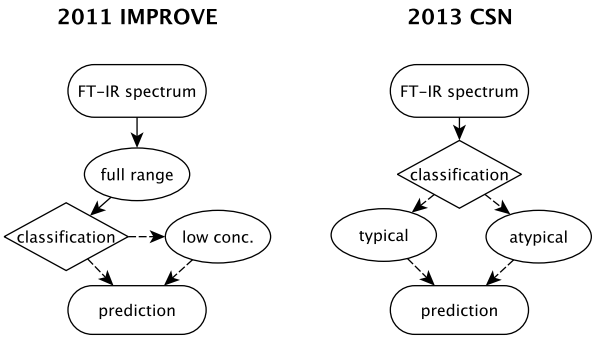

Figure 11. Multilevel modeling strategies used for TOR EC in the IMPROVE and CSN network. In the left figure, "full range" denotes the calibration model using the full range of TOR EC concentrations, while the "low conc." denotes the model using only the lowest third. In the right figure, "atypical" samples were taken from a particular site (ELLA) while the "typical" samples comprised the rest.

4 Operational phase of a calibration model

The operational phase of the model marks a departure from the building and evaluation phases (Figure 2) in that reference measurements may no longer be available on a regular basis. However, this is the eventual use case for such calibration models — for instance, to enable FT-IR to provide TOR-equivalent carbon values from a PTFE filter in new monitoring sites or measurement campaigns where TOR analysis from a separate filter is not available. Without reference measurements, it is important to evaluate the appropriateness of available calibration models for new samples, continually monitor the performance of the model by introspective means, and update the calibration as necessary ~~Within this constraint, we must continually monitor the performance of the model by introspective means, and update the calibration as necessary~~. To this end, we describe methods for anticipating prediction errors arising from precision and bias (Section 4.1), and strategies for calibration maintenance (Section 4.2).

4.1 Anticipating prediction errors for new samples

We dedicate this section to describe ways for anticipating prediction errors in new samples during the operational phase of a calibration model. Higher prediction errors may arise from a decrease in precision, or additional biases incurred for samples that are not well-represented by the calibration samples. The former can be approximated from the measurement noise characterized from the calibration set, while the latter is assessed on a more qualitative scale based on similarity of new samples to those in the calibration set. Anticipating these errors is imperative for reporting estimated precision for new samples, monitoring systematic changes in model performance, and selecting an alternate calibration model for new samples when prediction quality is questionable. For this task, we assume the unavailability of reference measurements for which evaluation methods in Sections 3.2.1 and 3.2.2 would otherwise apply; and primarily rely on spectral characteristics. To this end, Section 4.1.1 discusses the construction of prediction intervals around point estimates, Section 4.1.2 covers the strategy for outlier detection,

and Section 4.1.3 illustrates the use of sample similarity assessment for comparing suitability of models. The raw-spectra TOR EC calibration model for IMPROVE 2011 introduced by Dillner and Takahama (2015b) and evaluated for 2013 by Reggente et al. (2016) is revisited on account of its high prediction error and difficulty anticipating prediction errors compared to TOR OC.

4.1.1 Sample-specific prediction intervals

In Section 3, discussions focused around providing and evaluating point estimates of prediction. Additionally, interval estimates for each sample can be obtained to determine prediction uncertainty under a fixed relationship between model and data assumed under conditions of the calibration. In effect, prediction intervals describe magnitude of errors that are similar to those in the calibration set, and can be obtained from error propagation or resampling (bootstrap or jacknife) (Olivieri et al., 2006), or by employing a Bayesian framework (Murphy, 2012). We will restrict our discussion to estimating prediction intervals as they pertain to multivariate linear regression (including PLS). Provided that sufficient data exists, numerically-resampled intervals can be generated free of assumptions regarding underlying distributions, but the error propagation approach is favored on account of its connection to the fundamental processes contributing to the errors. The standard error of prediction has two primary contributions: the model contribution from calibration, and the measurement contribution from the prediction sample. These contribute nonlinearly to the prediction error, but an approximate expression can be derived through local linearization (i.e., neglecting higher-order terms typically assumed in error propagation) (Phatak et al., 1993; Denham, 1997; Faber et al., 2003; Serneels et al., 2004). This approximation results in a tractable expression for the prediction standard error $\sigma_{\hat{y},i}$ similar to that used by ordinary least squares regression, but considers heteroscedastic errors (Faber and Bro, 2002; ASTM E1655-17, 2017):

$$\sigma_{\hat{y},i} = s(1+h_i)^{1/2} \; . \tag{16}$$

The point estimate of prediction can then be bounded by an interval defined as $\pm t_{\alpha,\nu}\sigma_{\hat{y},i}$, where $t_{\alpha,\nu}$ denotes a t-distribution with significance level α and degrees of freedom ν. s is estimated from the fitting error — the mean squared error of calibration (MSEC: squared error normalized by the degrees of freedom). While a common assumption is that s captures only the prediction variance, the MSEC can implicitly include the prediction bias if present in the fit of the calibration set. h is the leverage introduced in eq. 15, and its role can be rationalized by the fact that samples closer to the "average" calibration sample are more precisely estimated than those which are further away. The approximations made for eq. 16 results in a method that is most applicable for small noise and small range of FT-IR absorbances (Faber and Kowalski, 1997a, b). Furthermore, prediction standard error can be refined by subtracting the precision of the reference measurement (Faber and Bro, 2002; Faber et al., 2003), but is not considered here.

The prediction intervals given by eq. 16 calculated for TOR-equivalent OC and EC are shown in Figure 12. Low standard errors of predictions anticipate low prediction errors, but prediction errors for higher concentrations (3–85 $\mu g\,cm^{-2}$) are more variable than indicated by the precision error. While deviations from observations in calibration are mostly explained by eq. 16, Reggente et al. (2016) and Weakley et al. (2018b) found that actual prediction errors do not always scale with computed

leverage. This phenomenon is also reported in other applications (Zhang and Garcia-Munoz, 2009), and indicates the possible role of bias due to differences in composition that is not well-captured by this metric.

It is also relevant to consider the standard errors of prediction for the TOR measurements (Chow et al., 2007a). Naïve propagation of reported errors across the relevant thermal fractions (including pyrolized carbon) leads to estimates of relative precision that approach 7 and 14% for TOR OC and EC, respectively, for the highest concentrations observed for this IMPROVE data set. As the errors are not truly independent for each sample, a simple summation of prediction variances may lead to an ~~overestimation~~underestimation. However, these calculated errors are close in magnitude to the average collocated precision error estimated for 2011 IMPROVE (15 and 23% for TOR OC and EC, respectively, Table 2), and the combined uncertainty estimated from analytical, cross-laboratory, and cross-sampler effects (Brown et al., 2017). The relative precision estimated for their respective calibration models using eq. 16 converges toward values which are approximately 3 times lower for both variables. The standard errors of prediction of a multivariate model can be lower than the reference measurements from which it is derived, as random errors from the latter are averaged out in the calibration process — especially when a large number of calibration samples are used (Difoggio, 1995). However, given that the apparent collocated precision for model predictions are on a par with TOR (Table 2), it is likely that model uncertainties calculated from eq. 16 are underestimated on account of unaccounted variations. Nonetheless, a general conclusion can still be drawn that many samples are predicted within uncertainty. There remain a samples (167 for TOR OC and 126 for TOR EC, out of 2177 total) that can be identified (in red, Figure 12) as having prediction errors which fall outside the anticipated range of uncertainty of both model and measurement. We describe procedures for algorithmically detecting these samples in the absence of reference measurements in Section 4.1.2.

4.1.2 Outlier detection

As described in Section 3.5, a calibration model that is likely to be suitable for a new sample is that which is trained on samples with similar concentration and composition. Therefore, identifying samples which are different from those in the calibration set of a particular model is closely tied to anticipation of potentially high prediction errors due to incurred bias. We first review possible categorizations of samples in a Venn diagram (Figure 13). Within a multivariate space encompassing all samples, some will lie at the edge of the domain (*extreme values*), while others will lie in sparsely populated regions of the interior (*inliers*). Some of these extreme values and inliers will be statistically surprising given the rest of the points, and are typically labeled as *outliers* or *anomalous samples* (Barnett and Lewis, 1994; Jouan-Rimbaud et al., 1999; Aggarwal, 2013). We note that inliers are sometimes used to refer to statistically different samples which lie within the composition domain, but we reserve the word outlier for all statistically significant samples in this paper. New samples in furthest proximity from calibration samples in this composition space require aggressive extrapolation or interpolation (i.e., they are least constrained by data), and are most likely to suffer in prediction performance. However, the actual increase in prediction error (if any) will depend on the functional relationship among variables and how well they are represented by the model — e.g., a linear relationship modeled by a linear mapping may perform adequately in interpolation and extrapolation. For instance, samples with OM/OC and OC/EC composition and TOR OC concentrations out of range with respect to calibration samples were predicted without substantial increase in errors (Section 3.5.1). Therefore, not all outliers may be associated with high prediction errors.

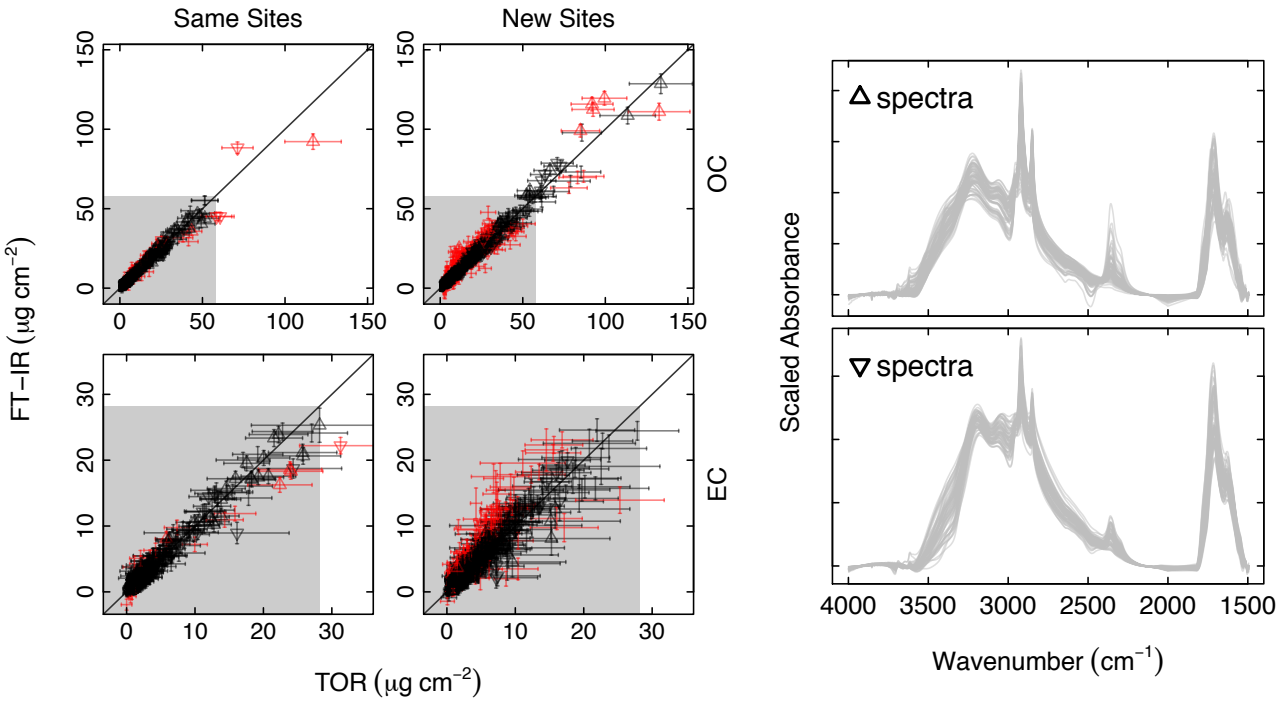

Figure 12. Point estimates and prediction intervals for the TOR-equivalent concentrations in the 2013 IMPROVE prediction set. Gray shades indicate extent of areal mass densities in the calibration samples. Triangles represent samples associated with wildfire burning (scaled spectra shown in right column). Red samples correspond to those for which the difference between predicted and observed concentrations exceed the combined uncertainties at the $\alpha = 0.05$ significance level.

Dissimilarity can be expressed as a measure of distance or a discrete label of normal/anomalous resulting from an unary (one-class) classification (Brereton, 2011). Identification of dissimilar observations is the subject of many disciplines including chemometrics, machine learning, and statistical process control and are referred under various names: anomaly detection, fault detection, novelty detection, and outlier detection (e.g. Wise and Gallagher, 1996; Montgomery, 2013; Pimentel et al., 2014). Together with knowledge regarding "prediction outliers" (samples with surprisingly high prediction errors), decisions can be grouped into the following outcomes (Figure 13): True Negative (TN; samples are classified as being similar and prediction error is low), True Positive (TP; samples are classified as being dissimilar and prediction error is high), False Negative (FN; samples are classified as being similar while prediction error is high), and False Positive (FP; samples are classified as being dissimilar while prediction error is low). The realization of these outcomes by a classifier can be used to judge its performance. We note that in contrast to the multilevel modeling strategy described in Section 3.5.3, the problem of error anticipation is to build a classifier that identifies all samples not similar to those in the training set (i.e., outliers, some of which may have

anomalously high magnitude of prediction error) without exhaustive knowledge or separate training sets comprising the new sample types.

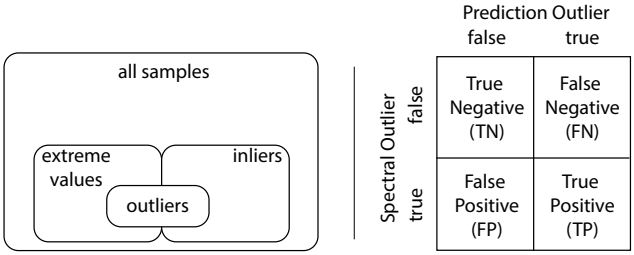

Figure 13. Venn diagram (not to scale), left, and confusion matrix, right, depicting the relationship between detected outliers and magnitude of prediction errors.

Without reference measurements, many external indicators might be used to characterize differences between new samples and those in the calibration set, especially with respect to attributes identified to be important (Section 3.5.1). For instance, the fraction of inorganic to total PM may give an indication of ammonium to OC ratio, or NO_x may be a valid surrogate for EC in many urban situations. However, our primary objective is to rely on indicators of composition and concentration that can be extracted directly from the FT-IR spectrum to determine the appropriateness of an existing calibration model to the new samples. Baseline corrected spectra have been used in the past to characterize similarity among ambient aerosol spectra through cluster analysis (e.g., Takahama et al., 2011; Ruthenburg et al., 2014), and can also be used for classification (Fearn, 2006; Isaksson and Aastveit, 2006). For instance, many of samples with large deviations in predictions of TOR-equivalent OC from observed values are spectroscopically similar (Figure 12) and exhibit sharp methylene peaks and large carbonyl absorbances present in spectra of biomass burning samples (Hawkins and Russell, 2010; Russell et al., 2011). Locations and dates of some these samples are consistent with known periods of wildfires, and will be the topic of future investigation. The underrepresentation of these types of samples in the 2011 IMPROVE calibration (and test) sets, or simply the higher concentrations beyond the calibration range may explain the proportionally high prediction errors incurred for these samples. The highest TOR EC concentrations in 2013 are associated with FRES, an urban site, and BYIS, an international site, both of which were not part of the 2011 calibration set. Spectral matching combined with model interpretation (Section 3.4) can identify particular sample types that may be problematic for a calibration model a priori. However, as sparse calibration modeling has shown (Section 3.3.2), not all spectral features are likely to be relevant for prediction of TOR OC or EC concentrations. Therefore, transformations specific for the target analyte (which can include but are not limited to spectral processing techniques described in Section 3.3) are likely to reveal the discriminating spectral features for distinguishing samples that are different from those in the calibration set.

Projection of the spectra in the feature space of the calibration model (i.e., factor scores and residuals of PLS or PCA, kernel distances, latent encoding in Gaussian process) after appropriate spectra processing and wavenumber selection can provide spectral comparisons that are specifically meaningful for prediction of the response variable (Nomikos and MacGregor, 1995;

MacGregor and Kourti, 1995; Camci et al., 2008; Ge and Song, 2010; Serradilla et al., 2011). For PLS regression, the feature vectors (scores) can be combined into a single metric called the Mahalanobis distance (Mahalanobis, 1936) or Hotelling's T^2 statistic (Hotelling, 1931), which are both proportional to the leverage introduced in eq. 15. The two terms are often used synonymously (e.g., Kourti and MacGregor, 1995; ASTM E1655-17, 2017), but can also be defined differently according to rank approximation of $\boldsymbol{X}$ or a coefficient making the T^2 comparable to the F-distribution (e.g. De Maesschalck et al., 2000; Brereton and Lloyd, 2016; Brereton, 2016). We will adopt the convention of defining $T^2 \equiv D_M^2$, but reserve Hotelling's T^2 statistic for use with its eponymous test to determine out-of-limit samples (e.g., in statistical process control) and D_M^2 for a general distance measure (which is also used in classification methods built upon different criteria). Outside of this feature space, the $Q^{(X)}$-statistic estimated using residuals $\boldsymbol{E}$ of spectra reconstructed from its latent variables (eq. 7) (Jackson, 2004) can additionally indicate variations orthogonal to the feature space, and hence variations which are orthogonal to the modeled portion of the response variable (Höskuldsson, 1996; Bro and Eldén, 2009). Therefore, $Q^{(X)}$ is typically monitored over time alongside T^2. The two metrics for mean-centered PLS can be written as follows:

$$T_i^2 = D_{M,i}^2 = (N-1) \cdot h$$
$$Q_i^{(X)} = \boldsymbol{e}_{X,i}\boldsymbol{e}_{X,i}^T = \boldsymbol{x}_i \left(\boldsymbol{I} - \boldsymbol{P}\boldsymbol{P}^T\right) \boldsymbol{x}_i^T$$

N is the number of samples in the calibration and h is the leverage from eq. 15. $\boldsymbol{P}$ is the matrix of loadings (eq. 6) and $\boldsymbol{e}_X$ denotes the row vector of residuals associated with each sample (eq. 5), equivalent to the product of latent variables unused for calibration. In an analytical chemistry context, high values of T^2 result from extreme values or unusual combinations of the same chemical components as those in the calibration set, whereas introduction of new analytes or interferences that result in spectroscopic response lying outside of the modeled domain would be carried in the residuals (Wise and Roginski, 2015). In practice, the separation of unfamiliar contributions to the spectra is likely not as clean, particularly with respect to nonlinear phenomena (e.g., scattering) which can be spread over multiple factors, and the portion of the spectroscopic signal associated with new substances may not be entirely apportioned to the residuals.

For classification purposes, thresholds for T^2 and $Q^{(X)}$ are determined from the F distribution and χ-square distribution, respectively, at different significance levels (Kourti and MacGregor, 1995). Classification and dissimilarity characterization by T^2 for a given data set performs best when the points converge toward a multivariate normal distribution. Such a distribution becomes less representative of the data set when the problem increases to proportions of extremely high dimensionality, where points become sparsely dispersed throughout the vast composition space rather than clustered around a single centroid (Domingos, 2012). To alleviate this problem, it is useful to conceptualize different relationships of training data in the column space of $\boldsymbol{T}$ and $\boldsymbol{E}$ against which new samples are compared. This task can be fulfilled by unary ~~(one-class)~~ classifiers that learn patterns from the data without imposition of global structure (e.g., normality). These approaches may employ superposition of local potential or kernel density functions (Jouan-Rimbaud et al., 1999; Latecki et al., 2007), kernel methods (Schölkopf et al., 1999), or recursive partitioning of the chemical space (Liu et al., 2008) for detection of points separated from the from the remainder of the samples.

For the 2013 IMPROVE data set, Reggente et al. (2016) used the 2011 IMPROVE calibration models developed by Dillner and Takahama (2015a; 2015b) and applied the Mahalanobis distance metric. Heuristic thresholds for D_M^2 and the prediction error were determined as their respective maximum values in the 2011 IMPROVE test set for purposes of classification. The number of samples in 2013 which had prediction errors greater than the selected threshold was small for both TOR OC and EC — for paired samples above detection limit across 17 sites ~~(and analytical blanks)~~, only 36 out of 2189 (TOR OC) and 22 out of 2177 (TOR EC) samples (1–2% of total) were determined as having high-errors according to this criterion. The overall accuracy (fraction of TN and TP out of total) was high, with 98% for both TOR OC and EC. These numbers are enviable for any classifier but was largely aided by the low number of high-error samples, which resulted in high overall accuracy from a permissive D_M^2 threshold and a limited number of FP classifications. When considering prediction intervals of both prediction and reference measurement, some of these high prediction errors are within anticipated uncertainties of the samples, while a few anomalous samples with errors outside of the range of uncertainties occur with lower absolute prediction errors (Section 4.1.1 and Figure 12). Therefore, we first correlate the results of outlier analysis to samples with prediction errors that lie outside of expected agreement (i.e., prediction outliers). We then revisit the topic of using these classification algorithms to identify samples with the highest magnitude of prediction errors.

For this discussion, it is useful to define two additional metrics: True Positive Rate (TPR) is the fraction of samples with high error correctly identified as such, and the False Positive Rate (FPR) is the fraction of samples with low errors that are incorrectly identified as having high error. In a coordinate space with TPR as the ordinate and FPR as the abscissa (Figure 14), the perfect model lies at (0, 1). For detecting new or anomalous spectra, we explore classifiers introduced above (potential function method, one-class SVM, and isolation forest) and consider their tradeoffs in TPR, FPR, and overall accuracy. For the potential function method, the radial basis function (RBF) is selected; the free parameters are the number of nearest neighbors used to determine the kernel width parameter and the confidence level for the thresholds. For one-class SVM, the RBF kernel is also used with the kernel coefficient and effective thresholding parameter varied. For isolation forest, the randomization seed and number of iterations is varied. For any given model, parameters or effective thresholds determine an approximate envelope in the space of TPR and FPR referred to as a Receiver Operating Characteristic (ROC) curve (Fawcett, 2006). For simplicity, the solutions with highest accuracy (fewest false classifications) and nearest proximity to the $(0,1)$ coordinate is shown in Figure 14, alongside T^2 and $Q^{(X)}$ for the $\alpha = \{0.01, 0.05, 0.1\}$ significance levels. For reference, the heuristic threshold for T^2 from Reggente et al. (2016) is also shown.

For TOR OC, classification performance using residuals ($\boldsymbol{E}$) is slightly but consistenty better than than using LVs ($\boldsymbol{T}$). The TPR ranges between 10–88% and FPR between 1–36% using $\boldsymbol{T}$ and TPR ranges between 36–87% and FPR between 4–28% using $\boldsymbol{E}$. For TOR EC, the selected results are clustered together with a few exceptions; TPRs and FPRs are typically higher (56–85% and 8–38%, respectively). Regarding systematic differences between methods over parameters studied, the potential function and SVM methods can span a wide range of solutions in the ROC space that follows the arc delineated by the selected points shown (up to TPR and FPR of 100%), while all isolation forest solutions remained in close proximity to the points depicted in Figure 14. Both T^2 and $Q^{(X)}$ metrics with the significance levels explored are restricted to the upper left corner of the ROC space as depicted.

The tradeoff in TPR and FPR is in part determined by what are designated as prediction outliers. The stratification of prediction errors by classification is illustrated in Figure 15. A classifier that is able to identify all samples with prediction errors greater than expected uncertainties would result in segregation by color in this figure. However, we see that the prediction outliers are only partially correlated with the absolute magnitude of prediction error (especially for TOR EC, where the pyrolized fraction adds a variable contribution to precision error across samples), while samples labeled as spectroscopic outliers are more aligned with the latter. Furthermore, samples with the lowest prediction errors are also not flagged as outliers. That ~~classifications are~~ spectral outliers are primarily correlated with magnitude of prediction errors ~~(without consideration for precision)~~ (more than deviation outside of expected precision) is ~~not surprising, as measurements with higher uncertainties have higher possibility of divergence from predictions modeled by the FT-IR spectra~~ sensible. Greater prediction errors are anticipated by sample leverage (eqs. 15 and 16) used explicitly or implicitly by classification algorithms, and high leverage can be related to extreme concentrations for which heteroscedastic measurement errors are also greater. Biomass burning samples previously mentioned can be identified visually (and by spectral matching), but they are not necessarily flagged as outliers with respect to the calibration models. This is not surprising as prediction errors for burning samples are not systematically higher, except for the few samples with highest TOR OC loadings. Revisiting the classification problem posed by Reggente et al. (2016) and considering only the samples with highest prediction errors exceeding those of the 2011 IMPROVE test set as prediction outliers, it is possible to achieve TPR of 81% and FPR of 12% for TOR OC, and TPR of 91% and FPR of 8% for TOR EC (both with the potential function method) as the solutions closest to $(0,1)$ on the ROC curve. Outlier detection for TOR EC is better served by alternative methods to T^2 on account of the strong non-normality in the multivariate feature space (Reggente et al., 2016). For this scenario, selecting a classifier with high TPR comes at a cost of lowering the overall accuracy significantly because of the small proportion of high-error samples. For instance, moving from the max D_M^2 classifier of Reggente et al. (2016) to the potential function solution for TOR EC as described above, an increase in TPR from 59% to 91% (a difference of 7 samples) accompanied by an increase in FPR from 1% to 8% (a difference of 142 samples) drops the overall accuracy from 98% to 92% on account of the large number of low-error samples that would be detected as being different. The desired criterion for the optimal classifier may depend on the purpose of classification. For the purposes of flagging suspicious samples during routine application of a calibration model, it may be desirable to select a classifier with high overall accuracy to keep the total number of FN and FP to a minimum. A conservative classifier with higher TPR than low FPR is, however, likely to be more useful for model selection against a specific sample (Section 4.1.3).

4.1.3 Model selection without reference measurements

Methods for error anticipation may also be used for evaluating among a set of candidate models when reference measurements are not available to provide a full evaluation. To illustrate such an application, we revisit the apparent increase in mean prediction error shown for decreasing number of ambient samples in the calibration set displayed in Figure 9. The corresponding increase in mean squared Mahalanobis distance between the fixed set of 253 test set spectra and those of the changing calibration set is shown in Figure 16. As D_M^2 increases linearly with the number of components, only the first 10 LVs are considered in each model for the purpose of a fair comparison. This example provides indication that the loss in representativeness of

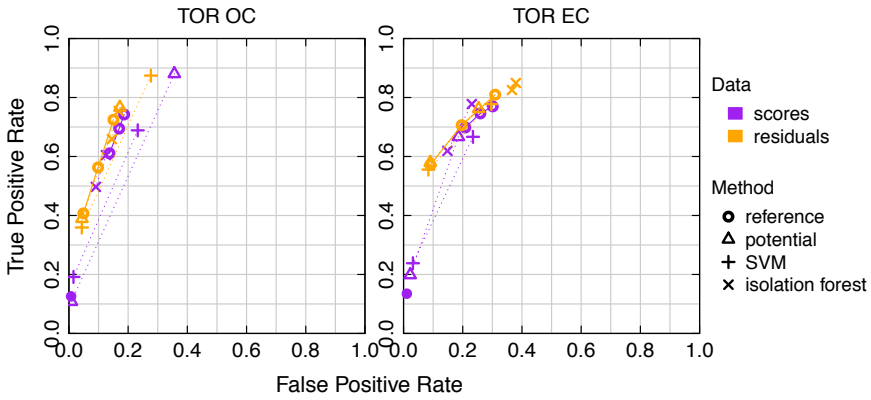

Figure 14. Receiver Operating Characteristic (ROC) curves for the 2013 IMPROVE data set. Symbol colors are grouped according to the data used for detection (either scores $\boldsymbol{T}$ or residuals $\boldsymbol{E}$). Symbol shapes indicate method of estimation. "Reference" denotes Hotelling's T^2 statistic for scores and the $Q^{(X)}$ statistic for residuals, for which three open circles are shown for the $\alpha = \{0.1, 0.05, 0.01\}$ significance levels. The filled purple symbol indicates the performance determined by the maximum T^2 of the 2011 IMPROVE test set, as originally used by Reggente et al. (2016). For other methods, two symbols are drawn and connected by dotted lines to indicate the solution with highest accuracy (fraction classified correctly) and the solution which lies closely to the coordinate $(0, 1)$.

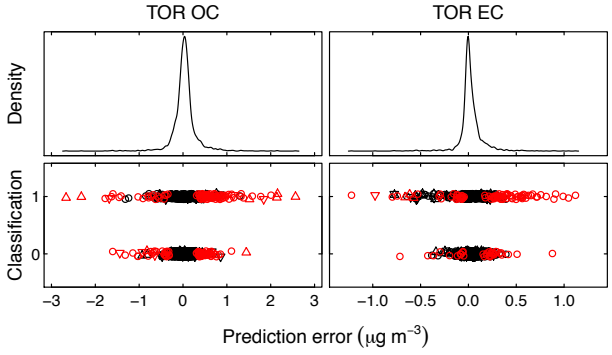

Figure 15. Prediction error distribution (top row) and classification results using the $Q^{(X)}$ classifier with $\alpha = 0.05$ significance level applied to model residuals (bottom row) for the 2013 IMPROVE data set. 1 corresponds to outliers and 0 as those not classified as outliers. Triangles and red samples correspond to same sample specification as Figure 12; rest of the individual prediction errors are symbolized with open circles.

composition or concentration between the 253 predicted samples and calibration samples as the latter numbers are diminished (Figure 10) is reflected in the FT-IR spectra, and can be appropriately extracted after projecting them onto factor scores of their respective PLS models.

While we have demonstrated use of D_M^2 to provide a qualitative comparison among several models, in principle it would be possible to use the classifiers introduced in Section 4.1.2 to find a set of models for which a new sample is not determined to be dissimilar. As mentioned in Section 4.1.2, a conservative classifier with higher TPR than low FPR is likely to be more useful for model selection for any specific sample. A sample-specific calibration model in which individual compounds from an available database for each new prediction sample is in principle possible using concepts described in this section. However, without a priori knowledge, the most relevant features and measure of similarity among individual samples is necessarily defined through the process of calibrating a model. Therefore, it is at present time necessary to hypothesize or propose several candidate models and select among them for any new prediction sample or set of samples for possible improvements in prediction.

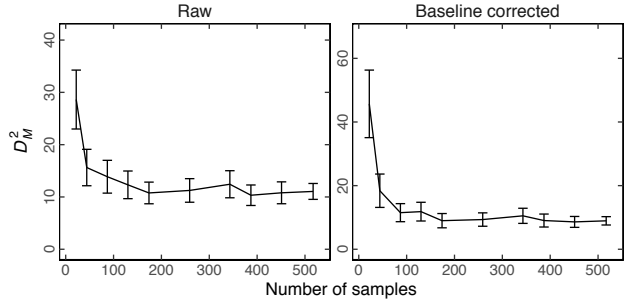

Figure 16. Mean squared Mahalanobis distance (D_M^2) between spectra of the fixed test set and changing calibration set, constructed as described in caption of Figure 9). Error bars span $\pm$ one standard deviation. The first 10 latent variables are used for estimation of D_M^2 in this example to reduce the dimensionality the factor space (Brereton and Lloyd, 2016).

4.2 ~~Calibration maintenance~~ Updating the calibration model

Calibration maintenance and transfer learning addresses the problem of updating a calibration model developed under one set of conditions to continue providing accurate predictions for samples measured under new conditions (Feudale et al., 2002; Torrey and Shavlik, 2009; Pan and Yang, 2010; Wise and Roginski, 2015). This topic has not yet been addressed for TOR OC and EC calibrations using FT-IR, but we can nonetheless make a few remarks for future research needs. Difference in sampled or measured conditions can arise from changes in hardware, changes in (PTFE filter) substrate, or atmospheric aerosol composition, and imply a possible difference introduced into distributions between training and prediction data in the feature space of the model. During the operational phase of the calibration, it is therefore necessary to continuously monitor model performance and appropriateness for new samples using protocols described in Section 3.2 and Section 4.1. Notable changes may be registered by trends in the magnitude of prediction errors compared against available reference measurements, or increasing instances of spectral outliers. The role of hardware performance in these changes can be assessed separately using the analytical protocols summarized in Section 2.3 — specifically, through the repeated analysis of laboratory check standards.

The strategy for model updating can be different according to the cause and nature of the change, but a basic premise is that the original condition still holds useful information that can be transferred to the new condition such that an entirely new

calibration is not warranted. In this way, a significant investment of resources required by model building (consisting of data collection and evaluation) may be avoided. For changes in instrument performance or installation of a separate spectrometer, commonly applied modifications range from simple linear corrections of predictions to calibration transfer algorithms to convert spectra to resemble that which may have been acquired from the primary instrument in its original state so that the original model remains applicable (Wise and Roginski, 2015; Chen et al., 2016b; Malli et al., 2017). The contribution from PTFE can presumably be removed with the appropriate baseline correction technique (Section 3.3.2). Though not been tested extensively across various filter types, successful prediction has been reported between two PTFE filter types (Weakley et al., 2018a). Treating the PTFE signal as an interferent, training the model with additional blank (zero-analyte) samples from different filter types may be an effective approach (Ottaway et al., 2012; Kalivas, 2012; Wise and Roginski, 2015), though also requires evaluation. Changing atmospheric composition can be addressed by updating the calibration set with new samples which contain new analytes or different regimes in concentration. While there are recursive algorithms for online updating (reweighting) of models with new samples (Hayes, 1996; Helland et al., 1992; Qin, 1998; Binfeng and Haibo, 2015; Ma et al., 2015; Chen et al., 2016b), recalibration with the appropriate proportion of old and new samples will recreate a feature space that accommodates both groups of samples. When new samples are needed, active learning strategies seek the potentially most informative samples and minimize the requirement of new calibration samples (Douak et al., 2012).

Additional strategies from *transductive learning* aim to avoid the requirement of obtaining new samples for recalibration, but rather search for common feature representations between calibration and prediction set ("unlabeled") samples (Chapelle et al., 2010). While these methods are more typically based on non-PLS based algorithms and applied to classification problems (Zadrozny, 2004; Cortes et al., 2005; Arnold et al., 2007; Bickel et al., 2007), some results in multivariate calibration tasks give an indication of their applicability. One approach is to reattribute weights in calibration samples to have the closest feature distribution to new samples (Huang et al., 2006; Sugiyama et al., 2008; Kim et al., 2011; Hazama and Kano, 2015; Zhang et al., 2017). New estimates weighted by their uncertainty can be furthermore be used for re-estimation of model parameters in an iterative fashion (Culp and Michailidis, 2008; Marcou et al., 2017). Another approach is to re-estimate a feature representation in which the calibration and prediction samples are in closer proximity in this space (Culp and Michailidis, 2008; Gujral et al., 2011; Pan et al., 2011). Limited studies with PLS regression report mixed results regarding the value of incorporating unlabeled data into the calibration over simply using the original model (Culp and Michailidis, 2008; Gujral et al., 2011; Paiva et al., 2012; Bao et al., 2015). The benefit of such efforts not surprisingly depend on both the specific characteristics of the calibration model and unlabeled data (Culp and Michailidis, 2008).

In the context of FT-IR measurements, TOR reference measurements may not be available for short-term campaigns at new sites and some aspects of transfer learning and transductive learning strategies (sample reweighting or basis-set rederivation) may be the only option for improvement if prediction errors from existing calibration models are expected to be high (Section 4.1.2). For long-term operation at a fixed site, collecting a limited number of reference samples for recalibration initially or periodically can be a viable strategy if sample characteristics substantially differ from those available for calibration. For instance, Reggente et al. (2016) showed that a recalibration strategy can improve predictions for new types of samples for the IMPROVE network. TOR predictions for samples collected in 2013 from FRES and BYIS sites had not only high instances

Table 3. Figures of merit for selected FRES (Fresno, CA) and BYIS (Baengnyeong Island, S. Korea) samples using base case 2011 IMPROVE calibration and a "dedicated" model built only using samples from FRES and BYIS.

Model	Variable	Samples	Bias ($\mu g\,m^{-3}$)	Error ($\mu g\,m^{-3}$)	R^2
2011 IMPROVE	OC	FRES, BYIS	0.28	0.43	0.79
Dedicated	OC	FRES, BYIS	-0.03	0.16	0.96
2011 IMPROVE	EC	FRES	0.05	0.10	0.85
Dedicated	EC	FRES	0	0.06	0.93
2011 IMPROVE	EC	BYIS	0.13	0.17	0.60
Dedicated	EC	BYIS	-0.07	0.11	0.66 / 0.84[*]

[*] one outlier removed.

of prediction errors, but also systematic biases when using the 2011 IMPROVE model. A dedicated calibration model built with two-thirds of the available data set at the two new sites improved prediction performance for samples reserved for testing (Table 3). Whether to incorporate new types of samples into the original calibration set to build a monolithic model, or to unify the calibrations through a multilevel modeling framework may depend on the number and leverage of new samples. A model derived from including new samples with old may cease to perform adequately for the original types of samples. From a case study in 2013 CSN (Weakley et al., 2018b), including ELLA samples in the calibration did not seem to affect the non-ELLA samples, but ELLA samples were also found to have not have much leverage within the scope of all samples. When updating an existing model, it is necessary to re-evaluate the model for old as well as new types of samples.

5 Conclusions

The FT-IR spectra of PM is rich in chemical information, and quantitative information such as TOR-equivalent OC and EC can be extracted from it provided that we can find the appropriate combination of training samples and algorithms for extraction. In this manuscript, we review procedures for spectral processing and data-driven calibration, where the data are taken from collocated measurements of TOR OC and EC. In this effort, procedures for initial steps for model building and evaluation, and later steps for monitoring of model behavior during the operational phase of a calibration model are described.

The number and types of samples required for calibration is determined by the diversity of composition in the prediction set. When samples are selected from the same sites as the prediction set, FT-IR calibration models could predict with virtually no bias and errors within 0.15 $\mu g\,m^{-3}$ for TOR OC and 0.11 $\mu g\,m^{-3}$ for TOR EC for areal loadings in the 2011 IMPROVE and 2013 CSN networks. Less than 5% of samples fell below the estimated detection limit. These metrics are on a par with the reference measurement evaluated for the same year. For the 2011 IMPROVE data set, the number of ambient calibration samples can be reduced from the canonical number of 501 down to approximately 150 samples and maintain similar prediction performance for the diversity in composition represented by 237 samples. To the extent that we have experimented (virtually) for TOR OC, the limitation is likely due to the difficulty in maintaining the same distribution of ammonium to OC ratio in the

calibration set as in the test set with fewer number of samples obtained by the temporal and spatial stratified sample reduction approach illustrated.

As evaluated for the IMPROVE network, TOR-equivalent concentrations in new samples collected for a later year (2013) and more sites (11 additional ones) have similar performance metrics overall, with exception to samples from two new sites (FRES and BYIS) not in the calibration set. Higher prediction errors for TOR OC occur largely due to specific types of samples not well-represented in the calibration year. While these samples are predicted without bias, their errors are higher on account of the higher areal loadings of TOR OC beyond the range of original calibration. Estimates of prediction intervals for both TOR and model predictions suggest that more than 92% of samples are predicted within anticipated precision errors. Outlier detection methods can be used to detect samples which are different with respect to the modeled domain to provide some indication of the magnitude of prediction errors. However, accurate detection of high-error samples comes with a tradeoff of increased false positive rates; the outlier detection method can be selected based on the application and desired tolerance for each type of detection error (false positive or negative). An obvious solution for reducing prediction errors in different samples is to acquire new samples for recalibration, though judicious calibration maintenance strategies (e.g., sample reweighting) can potentially minimize the number of new samples needed.

The procedure for quantitative prediction of TOR-equivalent OC and EC is a statistical one and depends the ability of an algorithm to resolve the overlapping absorption bands in the mid-IR and relate relevant features to the concentration of the target analyte. Given the evolving diversity in aerosol composition, it is not clear that arriving at an invariant, universal calibration model applicable for every new sample is practical. However, in describing the broader context of chemometrics and machine learning algorithms that are available for addressing each stage of the model life cycle, challenges for calibrating complex spectra are not insurmountable provided that they are systematically handled as described in this paper. We can use a wide range of statistical quality control procedures at our disposal to assess similarity of relevant features among spectra to continually monitor model performance, to anticipate appropriateness of existing calibration models, and to propose revisions. Construction of calibration models specific to individual or groups of samples may be envisioned provided that we are further able to identify the most important spectral features to assess similarities relevant for TOR OC and EC estimation.

In parallel to ensuring numerical accuracy of a calibration, understanding how the calibration relates spectral absorbances to TOR concentrations is critical for anticipating model applicability. Identification of important vibrational modes used in the calibration facilitates understanding of how the model relates absorbances to concentrations of the target analyte. Moreover, this association can be used to gain a better understanding of molecular structure in complex substances underlying the OC and EC concentrations reported by TOR. For TOR-equivalent OC, functional groups typically associated with atmospheric organic matter were found: aliphatic CH, carbonyls, and nitrogenated functional groups. For TOR-equivalent EC prediction, the vibrational mode associated with C-C stretch of aromatic rings typically observed in mid-IR spectra of soot appears to be an important absorption band, but a model for Elizabeth, NJ, was able to predict TOR-equivalent EC concentrations accurately without use of this spectroscopic region. While attempts to understand model LVs have thus far been limited, some work by Weakley et al. (2016) indicate that 2013 CSN aerosols could be modeled with a surprisingly few LVs, with nearly 90% of the

variation in TOR OC explained by one variable. Further analysis of constituent samples using source apportionment techniques and analysis of chemical composition (e.g., using functional groups) are bound to benefit overall model interpretation.

In summary, this manuscript outlines a general perspective and specific practices for model building; encompassing judicious specification of algorithm, spectra processing procedure, and sample selection. Taking a systematic approach toward calibration with a diverse set of reference measurements allows us to expand the suite of information extractable from FT-IR spectra, to complement functional group analysis from laboratory calibrations, which has long been the focus. Given the demonstrated simplicity and non-destructive nature of acquiring spectra from PTFE filters, this technique can expand TOR-equivalent OC and EC measurements (which has a long history) to new campaigns and new locations in which only PTFE samples are collected for gravimetric reference measurements. Therefore, we anticipate that the procedure outlined in this paper can complement existing methods for PM monitoring with TOR-equivalent OC and EC, and provide guidance in extracting composition of substances from FT-IR spectra of atmospheric PM. Given that a wide range of inorganic and organic substances display mid-IR activity, further exploration of data sources and algorithms for quantitative analysis can continue to expand the cost-effective application of FT-IR in chemical speciation measurements.

Code availability. Companion paper and source code is under review and its functionality is made accesible through the web platform http://airspec.epfl.ch.

Data availability. The IMPROVE and CSN network data will be made publically available.

Competing interests. The authors have no competing interests.

Disclaimer. None.

Acknowledgements. The authors acknowledge funding from EPFL, Swiss National Science Foundation (200021_143298, 200021_169506), Electric Power Research Institute (MA10003745), and the U.S. EPA and the IMPROVE program (National Park Service cooperative agreement P11AC91045). We also thank the IMPROVE team at UC Davis for performing the sample handling and site maintenance for all IMPROVE sites and the RTI International team for managing the CSN during the 2013 sampling year.

References

Abdi, H.: Partial least squares regression and projection on latent structure regression (PLS Regression), *Wiley Interdisciplinary Reviews: Computational Statistics*, 2, 97–106, https://doi.org/10.1002/wics.51, 2010.

Afseth, N. K. and Kohler, A.: Extended multiplicative signal correction in vibrational spectroscopy, a tutorial, *Chemometrics and Intelligent Laboratory Systems*, 117, 92–99, https://doi.org/10.1016/j.chemolab.2012.03.004, 2012.

Aggarwal, C. C.: Outlier Analysis, Springer Publishing Company, Incorporated, 2013.

Aida, M. and Dupuis, M.: IR and Raman intensities in vibrational spectra from direct ab initio molecular dynamics: D2O as an illustration, *Journal of Molecular Structure: THEOCHEM*, 633, 247–255, https://doi.org/10.1016/S0166-1280(03)00280-X, 2003.

Aitken, A. C.: IV.—On Least Squares and Linear Combination of Observations, *Proceedings of the Royal Society of Edinburgh*, 55, 42–48, https://doi.org/10.1017/S0370164600014346, 1936.

Akaike, H.: A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, 19, 716–723, https://doi.org/10.1109/TAC.1974.1100705, 1974.

Akhter, M. S., Chughtai, a. R., and Smith, D. M.: The Structure of Hexane Soot I: Spectroscopic Studies, *Applied Spectroscopy*, 39, 143–153, https://doi.org/10.1366/0003702854249114, 1985.

Akimoto, H., Bandow, H., Sakamaki, F., Inoue, G., Hoshino, M., and Okuda, M.: Photooxidation of the propylene-NOx-air system studied by long-path Fourier transform infrared spectrometry, *Environmental Science & Technology*, 14, 172–179, https://doi.org/10.1021/es60162a007, 1980.

Allen, D. T. and Palen, E.: Recent advances in aerosol analysis by infrared spectroscopy, *Journal of Aerosol Science*, 20, 441–455, https://doi.org/10.1016/0021-8502(89)90078-5, 1989.

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.

Andries, E. and Kalivas, J. H.: Interrelationships between generalized Tikhonov regularization, generalized net analyte signal, and generalized least squares for desensitizing a multivariate calibration to interferences, *Journal of Chemometrics*, 27, 126–140, https://doi.org/10.1002/cem.2501, 2013.

Arlot, S. and Celisse, A.: A survey of cross-validation procedures for model selection, *Statistics Surveys*, 4, 40–79, https://doi.org/10.1214/09-SS054, 2010.

Arnold, A., Nallapati, R., and Cohen, W. W.: A Comparative Study of Methods for Transductive Transfer Learning, in: Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007), pp. 77–82, https://doi.org/10.1109/ICDMW.2007.109, 2007.

ASTM D7844-12: Standard Test Method for Condition Monitoring of Soot in In-Service Lubricants by Trend Analysis using Fourier Transform Infrared (FT-IR) Spectrometry, Standard D7844-12, West Conshohocken, PA, https://doi.org/10.1520/D7844-12, http://www.astm.org, 2017.

ASTM E1655-17: Standard Practices for Infrared Multivariate Quantitative Analysis, Standard E1655-17, West Conshohocken, PA, https://doi.org/10.1520/E1655-17, http://www.astm.org, 2017.

Balabin, R. M. and Smirnov, S. V.: Variable selection in near-infrared spectroscopy: Benchmarking of feature selection methods on biodiesel data, *Analytica Chimica Acta*, 692, 63–72, 2011.

Bao, L., Yuan, X., and Ge, Z.: Co-training partial least squares model for semi-supervised soft sensor development, *Chemometrics and Intelligent Laboratory Systems*, 147, 75–85, https://doi.org/10.1016/j.chemolab.2015.08.002, 2015.

Barnett, V. and Lewis, T.: Outliers in Statistical Data, Wiley Series in Probability & Statistics, Wiley, 1994.

Barone, V., Baiardi, A., Biczysko, M., Bloino, J., Cappelli, C., and Lipparini, F.: Implementation and validation of a multi-purpose virtual spectrometer for large systems in complex environments, *Physical Chemistry Chemical Physics*, 14, 12 404–12 422, https://doi.org/10.1039/C2CP41006K, 2012.

Barone, V., Biczysko, M., and Bloino, J.: Fully anharmonic IR and Raman spectra of medium-size molecular systems: accuracy and interpretation, *Physical Chemistry Chemical Physics*, 16, 1759–1787, https://doi.org/10.1039/C3CP53413H, 2014.

Barth, A.: SpecInfo: An integrated spectroscopic information system, *Journal of Chemical Information and Computer Sciences*, 33, 52–58, https://doi.org/10.1021/ci00011a009, 1993.

Baumann, K. and Clerc, J. T.: Computer-assisted IR spectra prediction — linked similarity searches for structures and spectra, *Analytica Chimica Acta*, 348, 327–343, https://doi.org/10.1016/S0003-2670(97)00238-9, 1997.

Behler, J. and Parrinello, M.: Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces, *Physical Review Letters*, 98, 146 401, https://doi.org/10.1103/PhysRevLett.98.146401, 2007.

Bernasconi, M., Silvestrelli, P. L., and Parrinello, M.: Ab Initio Infrared Absorption Study of the Hydrogen-Bond Symmetrization in Ice, *Physical Review Letters*, 81, 1235–1238, https://doi.org/10.1103/PhysRevLett.81.1235, 1998.

Bickel, S., Brückner, M., and Scheffer, T.: Discriminative Learning for Differing Training and Test Distributions, in: Proceedings of the 24th International Conference on Machine Learning, ICML '07, pp. 81–88, ACM, New York, NY, USA, https://doi.org/10.1145/1273496.1273507, 2007.

Binfeng, Y. and Haibo, J.: Near-infrared calibration transfer via support vector machine and transfer learning, *Analytical Methods*, 7, 2714–2725, https://doi.org/10.1039/C4AY02462A, 2015.

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

Blando, J. D., Porcja, R. J., Li, T. H., Bowman, D., Lioy, P. J., and Turpin, B. J.: Secondary formation and the Smoky Mountain organic aerosol: An examination of aerosol polarity and functional group composition during SEAVS RID F-6148-2011, *Environmental Science & Technology*, 32, 604–613, https://doi.org/10.1021/es970405s, 1998.

Bogard, J. S., Johnson, S. A., Kumar, R., and Cunningham, P. T.: Quantitative analysis of nitrate ion in ambient aerosols by Fourier-transform infrared spectroscopy, *Environmental Science & Technology*, 16, 136–140, https://doi.org/10.1021/es00097a004, 1982.

Borggaard, C. and Thodberg, H. H.: Optimal minimal neural interpretation of spectra, *Analytical Chemistry*, 64, 545–551, https://doi.org/10.1021/ac00029a018, 1992.

Bornemann, L., Welp, G., Brodowski, S., Rodionov, A., and Amelung, W.: Rapid assessment of black carbon in soil organic matter using mid-infrared spectroscopy, *Organic Geochemistry*, 39, 1537–1544, https://doi.org/10.1016/j.orggeochem.2008.07.012, 2008.

Brereton, R. G.: One-class classifiers, *Journal of Chemometrics*, 25, 225–246, https://doi.org/10.1002/cem.1397, 2011.

Brereton, R. G.: Hotelling's T squared distribution, its relationship to the F distribution and its use in multivariate space, *Journal of Chemometrics*, 30, 18–21, https://doi.org/10.1002/cem.2763, 2016.

Brereton, R. G. and Lloyd, G. R.: Re-evaluating the role of the Mahalanobis distance measure, *Journal of Chemometrics*, 30, 134–143, https://doi.org/10.1002/cem.2779, 2016.

Bro, R. and Eldén, L.: PLS works, *Journal of Chemometrics*, 23, 69–71, https://doi.org/10.1002/cem.1177, 2009.

Brown, P. J., Fearn, T., and Vannucci, M.: Bayesian Wavelet Regression on Curves With Application to a Spectroscopic Calibration Problem, *Journal of the American Statistical Association*, 96, 398–408, https://doi.org/10.1198/016214501753168118, 2001.

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.

Burbidge, J. B., Magee, L., and Robb, A. L.: Alternative Transformations to Handle Extreme Values of the Dependent Variable, *Journal of the American Statistical Association*, 83, 123–127, 1988.

Cain, J. P., Gassman, P. L., Wang, H., and Laskin, A.: Micro-FTIR study of soot chemical composition-evidence of aliphatic hydrocarbons on nascent soot surfaces, *Physical Chemistry Chemical Physics*, 12, 5206–5218, https://doi.org/10.1039/b924344e, 2010.

Camci, F., Chinnam, R. B., and Ellis, R. D.: Robust kernel distance multivariate control chart using support vector principles, *International Journal of Production Research*, 46, 5075–5095, https://doi.org/10.1080/00207540500543265, 2008.

Cappelli, C. and Biczysko, M.: Time-Independent Approach to Vibrational Spectroscopies, in: Computational Strategies for Spectroscopy, edited by Barone, V., pp. 309–360, John Wiley & Sons, Inc., https://doi.org/10.1002/9781118008720.ch7/summary, 2011.

Car, R. and Parrinello, M.: Unified Approach for Molecular Dynamics and Density-Functional Theory, *Physical Review Letters*, 55, 2471–2474, https://doi.org/10.1103/PhysRevLett.55.2471, 1985.

Caruana, R.: Multitask Learning, *Machine Learning*, 28, 41–75, https://doi.org/10.1023/A:1007379606734, 1997.

Ceriotti, M., Fang, W., Kusalik, P. G., McKenzie, R. H., Michaelides, A., Morales, M. A., and Markland, T. E.: Nuclear Quantum Effects in Water and Aqueous Systems: Experiment, Theory, and Current Challenges, *Chemical Reviews*, 116, 7529–7550, https://doi.org/10.1021/acs.chemrev.5b00674, 2016.

Chapelle, O., Schölkopf, B., and Zien, A.: Semi-Supervised Learning, The MIT Press, 1st edn., 2010.

Chen, Q., Ikemori, F., Higo, H., Asakawa, D., and Mochida, M.: Chemical Structural Characteristics of HULIS and Other Fractionated Organic Matter in Urban Aerosols: Results from Mass Spectral and FT-IR Analysis, *Environmental Science & Technology*, 50, 1721–1730, https://doi.org/10.1021/acs.est.5b05277, 2016a.

Chen, T. and Yang, Y.: Interpretation of non-linear empirical data-based process models using global sensitivity analysis, *Chemometrics and Intelligent Laboratory Systems*, 107, 116–123, https://doi.org/10.1016/j.chemolab.2011.02.006, 2011.

Chen, T., Morris, J., and Martin, E.: Gaussian process regression for multivariate spectroscopic calibration, *Chemometrics and Intelligent Laboratory Systems*, 87, 59–71, 2007.

Chen, W.-R., Bin, J., Lu, H.-M., Zhang, Z.-M., and Liang, Y.-Z.: Calibration transfer via an extreme learning machine auto-encoder, *Analyst*, 141, 1973–1980, https://doi.org/10.1039/C5AN02243F, 2016b.

Cheng, C.-H., Lehmann, J., and Engelhard, M. H.: Natural oxidation of black carbon in soils: Changes in molecular form and surface charge along a climosequence, *Geochimica et Cosmochimica Acta*, 72, 1598–1610, https://doi.org/10.1016/j.gca.2008.01.010, 2008.

Chong, I. G. and Jun, C. H.: Performance of some variable selection methods when multicollinearity is present, *Chemometrics and Intelligent Laboratory Systems*, 78, 103–112, https://doi.org/10.1016/j.chemolab.2004.12.011, 2005.

Chow, J. C.: Measurement Methods to Determine Compliance with Ambient Air Quality Standards for Suspended Particles, *Journal of the Air & Waste Management Association*, 45, 320–382, https://doi.org/10.1080/10473289.1995.10467369, 1995.

Chow, J. C., Watson, J. G., Pritchett, L. C., Pierson, W. R., Frazier, C. A., and Purcell, R. G.: The dri thermal/optical reflectance carbon analysis system: description, evaluation and applications in U.S. Air quality studies, *Atmospheric Environment. Part A. General Topics*, 27, 1185–1201, https://doi.org/10.1016/0960-1686(93)90245-T, 1993.

Chow, J. C., Watson, J. G., Chen, L.-W. A., Arnott, W. P., Moosmüller, H., and Fung, K.: Equivalence of Elemental Carbon by Thermal/Optical Reflectance and Transmittance with Different Temperature Protocols, *Environmental Science & Technology*, 38, 4414–4422, https://doi.org/10.1021/es034936u, 2004.

Chow, J. C., Watson, J. G., Chen, L.-W. A., Chang, M. O., Robinson, N. F., Trimble, D., and Kohl, S.: The IMPROVE_A Temperature Protocol for Thermal/Optical Carbon Analysis: Maintaining Consistency with a Long-Term Database, *Journal of the Air & Waste Management Association*, 57, 1014–1023, https://doi.org/10.3155/1047-3289.57.9.1014, 2007a.

Chow, J. C., Yu, J. Z., Watson, J. G., Ho, S. S. H., Bohannan, T. L., Hays, M. D., and Fung, K. K.: The application of thermal methods for determining chemical composition of carbonaceous aerosols: A review, *Journal of Environmental Science and Health Part A-Toxic/Hazardous Substances & Environmental Engineering*, 42, 1521–1541, https://doi.org/10.1080/10934520701513365, 2007b.

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.

Christian, T. J., Kleiss, B., Yokelson, R. J., Holzinger, R., Crutzen, P. J., Hao, W. M., Shirai, T., and Blake, D. R.: Comprehensive laboratory measurements of biomass-burning emissions: 2. First intercomparison of open-path FTIR, PTR-MS, and GC-MS/FID/ECD, *Journal of Geophysical Research-Atmospheres*, 109, https://doi.org/10.1029/2003JD003874, 2004.

Christie, B. D. and Munk, M. E.: Structure generation by reduction: a new strategy for computer-assisted structure elucidation, *Journal of Chemical Information and Computer Sciences*, 28, 87–93, https://doi.org/10.1021/ci00058a009, 1988.

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.

Cortes, C., Mohri, M., and Weston, J.: A General Regression Technique for Learning Transductions, in: Proceedings of the 22Nd International Conference on Machine Learning, ICML '05, pp. 153–160, ACM, New York, NY, USA, https://doi.org/10.1145/1102351.1102371, 2005.

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.

Cross, E. S., Williams, L. R., Lewis, D. K., Magoon, G. R., Onasch, T. B., Kaminsky, M. L., Worsnop, D. R., and Jayne, J. T.: Use of electrochemical sensors for measurement of air pollution: correcting interference response and validating measurements, *Atmospheric Measurement Techniques*, 10, 3575–3588, https://doi.org/10.5194/amt-10-3575-2017, 2017.

Culp, M. and Michailidis, G.: An Iterative Algorithm for Extending Learners to a Semi-Supervised Setting, *Journal of Computational and Graphical Statistics*, 17, 545–571, https://doi.org/10.1198/106186008X344748, 2008.

Cunningham, P. T. and Johnson, S. A.: Spectroscopic observation of acid sulfate in atmospheric particulate samples, *Science*, 191, 77–79, https://doi.org/10.1126/science.1856, 1976.

Cunningham, P. T., Johnson, S. A., and Yang, R. T.: Variations in chemistry of airborne particulate material with particle size and time, *Environmental Science & Technology*, 8, 131–135, https://doi.org/10.1021/es60087a002, 1974.

Cziczo, D. J., Nowak, J. B., Hu, J. H., and Abbatt, J. P. D.: Infrared spectroscopy of model tropospheric aerosols as a function of relative humidity: Observation of deliquescence and crystallization, *Journal of Geophysical Research-atmospheres*, 102, 18 843–18 850, https://doi.org/10.1029/97JD01361, 1997.

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.

de Juan, A. and Tauler, R.: Multivariate Curve Resolution (MCR) from 2000: Progress in Concepts and Applications, *Critical Reviews in Analytical Chemistry*, 36, 163–176, https://doi.org/10.1080/10408340600970005, 2006.

De Maesschalck, R., Jouan-Rimbaud, D., and Massart, D.: The Mahalanobis distance, *Chemometrics and Intelligent Laboratory Systems*, 50, 1–18, https://doi.org/10.1016/S0169-7439(99)00047-7, 2000.

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*, accepted, https://doi.org/10.1177/0003702818804574, 2018.

Decesari, S., Facchini, M. C., Mircea, M., Cavalli, F., and Fuzzi, S.: Solubility properties of surfactants in atmospheric aerosol and cloud/fog water samples, *Journal of Geophysical Research-Atmospheres*, 108, 4685, https://doi.org/10.1029/2003JD003566, 2003.

deJong, S.: Simpls - An Alternative Approach To Partial Least-squares Regression, *Chemometrics and Intelligent Laboratory Systems*, 18, 251–263, https://doi.org/10.1016/0169-7439(93)85002-X, 1993.

Denham, M. C.: Prediction intervals in partial least squares, *J. Chemometrics*, 11, 39–52, 1997.

DeNoyer, L. and Dodd, J. G.: Smoothing and Derivatives in Spectroscopy, John Wiley & Sons, Ltd, https://doi.org/10.1002/0470027320.s4501, 2006.

Despagne, F. and Luc Massart, D.: Neural networks in multivariate calibration, *Analyst*, 123, 157–178, https://doi.org/10.1039/A805562I, 1998.

Difoggio, R.: Examination of Some Misconceptions about Near-Infrared Analysis, *Applied Spectroscopy*, 49, 67–75, https://doi.org/10.1366/0003702953963247, 1995.

Dillner, A. M.: Change to artifact correction method for OC carbon fractions, http://vista.cira.colostate.edu/improve/Data/QA_QC/Advisory/da0032/da0032_OC_artifact.pdf, accessed: 2018-02-18, 2018.

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, 2015a.

Dillner, A. M. and Takahama, S.: Predicting ambient aerosol thermal-optical reflectance measurements from infrared spectra: elemental carbon, *Atmospheric Measurement Techniques*, 8, 4013–4023, https://doi.org/10.5194/amt-8-4013-2015, 2015b.

Dodd, J. G. and DeNoyer, L.: Curve-Fitting: Modeling Spectra, John Wiley & Sons, Ltd, https://doi.org/10.1002/0470027320.s4503, 2006.

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.

Douak, F., Melgani, F., Alajlan, N., Pasolli, E., Bazi, Y., and Benoudjit, N.: Active learning for spectroscopic data regression, *Journal of Chemometrics*, 26, 374–383, https://doi.org/10.1002/cem.2443, 2012.

Doughty, D. C. and Hill, S. C.: Automated aerosol Raman spectrometer for semi-continuous sampling of atmospheric aerosol, *Journal of Quantitative Spectroscopy and Radiative Transfer*, 188, 103–117, https://doi.org/10.1016/j.jqsrt.2016.06.042, 2017.

Dubois, J. E., Mathieu, G., Peguet, P., Panaye, A., and Doucet, J. P.: Simulation of infrared spectra: an infrared spectral simulation program (SIRS) which uses DARC topological substructures, *Journal of Chemical Information and Computer Sciences*, 30, 290–302, https://doi.org/10.1021/ci00067a013, 1990.

Duyckaerts, G.: The infra-red analysis of solid substances. A review, *Analyst*, 84, 201–214, https://doi.org/10.1039/AN9598400201, 1959.

Efron, B. and Tibshirani, R.: Improvements on Cross-Validation: The .632+ Bootstrap Method, *Journal of the American Statistical Association*, 92, 548–560, 1997.

Eilers, P. H. C.: Parametric Time Warping, *Analytical Chemistry*, 76, 404–411, https://doi.org/10.1021/ac034800e, 2004.

Elyashberg, M., Blinov, K., Molodtsov, S., Smurnyy, Y., Williams, A. J., and Churanova, T.: Computer-assisted methods for molecular structure elucidation: realizing a spectroscopist's dream, *Journal of Cheminformatics*, 1, 3, https://doi.org/10.1186/1758-2946-1-3, 2009.

Esbensen, K. H. and Geladi, P.: Principles of Proper Validation: use and abuse of re-sampling for validation, *Journal of Chemometrics*, 24, Hungarian Chem Soc, https://doi.org/10.1002/cem.1310, 2010.

Faber, K. and Kowalski, B. R.: Propagation of measurement errors for the validation of predictions obtained by principal component regression and partial least squares, *Journal of Chemometrics*, 11, 181–238, https://doi.org/10.1002/(SICI)1099-128X(199705)11:3<181::AID-CEM459>3.0.CO;2-7, 1997a.

Faber, K. and Kowalski, B. R.: Improved prediction error estimates for multivariate calibration by correcting for the measurement error in the reference values, *Applied Spectroscopy*, 51, 660–665, https://doi.org/10.1366/0003702971941061 EOLEOL, 1997b.

Faber, N. K. M. and Bro, R.: Standard error of prediction for multiway PLS: 1. Background and a simulation study, *Chemometrics and Intelligent Laboratory Systems*, 61, 133–149, https://doi.org/10.1016/S0169-7439(01)00204-0, 2002.

Faber, N. M., Song, X. H., and Hopke, P. K.: Sample-specific standard error of prediction for partial least squares regression, *Trac-trends in Analytical Chemistry*, 22, 330–334, https://doi.org/10.1016/S0165-9936(03)00503-X, 2003.

Faber, P., Drewnick, F., Bierl, R., and Borrmann, S.: Complementary online aerosol mass spectrometry and offline FT-IR spectroscopy measurements: Prospects and challenges for the analysis of anthropogenic aerosol particle emissions, *Atmospheric Environment*, 166, 92–98, https://doi.org/10.1016/j.atmosenv.2017.07.014, 2017.

Farrés, M., Platikanov, S., Tsakovski, S., and Tauler, R.: Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation, *Journal of Chemometrics*, 29, 528–536, https://doi.org/10.1002/cem.2736, 2015.

Fawcett, T.: An introduction to ROC analysis, *Pattern Recognition Letters*, 27, 861–874, https://doi.org/10.1016/j.patrec.2005.10.010, 2006.

Fearn, T.: Discriminant Analysis, in: Handbook of Vibrational Spectroscopy, John Wiley & Sons, Ltd, https://doi.org/10.1002/0470027320.s4302, 2006.

Feudale, R. N., Woody, N. A., Tan, H., Myles, A. J., Brown, S. D., and Ferré, J.: Transfer of multivariate calibration models: a review, *Chemometrics and Intelligent Laboratory Systems*, 64, 181–192, https://doi.org/10.1016/S0169-7439(02)00085-0, 2002.

Filzmoser, P., Gschwandtner, M., and Todorov, V.: Review of sparse methods in regression and classification with application to chemometrics, *Journal of Chemometrics*, 26, 42–51, https://doi.org/10.1002/cem.1418, 2012.

Fischer, S., Ueltschi, T., El-Khoury, P., Mifflin, A., Hess, W., Wang, H., Cramer, C., and Govind, N.: Infrared and Raman Spectroscopy from Ab Initio Molecular Dynamics and Static Normal Mode Analysis: The C-H Region of DMSO as a Case Study, *Journal of Physical Chemistry B Materials*, 120, 1429–1436, https://doi.org/10.1021/acs.jpcb.5b03323, 2016.

Flores, E., Viallon, J., Moussay, P., and Wielgosz, R. I.: Accurate Fourier Transform Infrared (FT-IR) Spectroscopy Measurements of Nitrogen Dioxide (NO2) and Nitric Acid (HNO3) Calibrated with Synthetic Spectra, *Applied Spectroscopy*, 67, 1171–1178, https://doi.org/10.1366/13-07030, 2013.

Flores, E., Viallon, J., Moussay, P., Griffith, D. W. T., and Wielgosz, R. I.: Calibration Strategies for FT-IR and Other Isotope Ratio Infrared Spectrometer Instruments for Accurate $\delta 13$C and $\delta 18$O Measurements of CO2 in Air, *Analytical Chemistry*, 89, 3648–3655, https://doi.org/10.1021/acs.analchem.6b05063, 2017.

Foster, R. D. and Walker, R. F.: Quantitative determination of crystalline silica in respirable-size dust samples by infrared spectrophotometry, *Analyst*, 109, 1117–1127, https://doi.org/10.1039/AN9840901117, 1984.

Friedel, R. and Carlson, G.: Difficult carbonaceous materials and their infra-red and Raman spectra. Reassignments for coal spectra, *Fuel*, 51, 194–198, https://doi.org/10.1016/0016-2361(72)90079-8, 1972.

Friedel, R. A. and Carlson, G. L.: Infrared spectra of ground graphite, *The Journal of Physical Chemistry*, 75, 1149–1151, https://doi.org/10.1021/j100678a021, 1971.

Friedman, J. H., Hastie, T., and Tibshirani, R.: Regularization Paths for Generalized Linear Models via Coordinate Descent, *Journal of Statistical Software*, 33, 1–22, 2010.

Fu, G.-H., Xu, Q.-S., Li, H.-D., Cao, D.-S., and Liang, Y.-Z.: Elastic Net Grouping Variable Selection Combined with Partial Least Squares Regression (EN-PLSR) for the Analysis of Strongly Multi-collinear Spectroscopic Data, *Applied Spectroscopy*, 65, 402–408, https://doi.org/10.1366/10-06069, 2011.

Gaigeot, M.-P.: Alanine Polypeptide Structural Fingerprints at Room Temperature: What Can Be Gained from Non-Harmonic Car−Parrinello Molecular Dynamics Simulations, *The Journal of Physical Chemistry A*, 112, 13 507–13 517, https://doi.org/10.1021/jp807550j, 2008.

Gaigeot, M.-P. and Sprik, M.: Ab Initio Molecular Dynamics Computation of the Infrared Spectrum of Aqueous Uracil, *The Journal of Physical Chemistry B*, 107, 10 344–10 358, https://doi.org/10.1021/jp034788u, 2003.

Gaigeot, M.-P., Martinez, M., and Vuilleumier, R.: Infrared spectroscopy in the gas and liquid phase from first principle molecular dynamics simulations: application to small peptides, *Molecular Physics*, 105, 2857–2878, https://doi.org/10.1080/00268970701724974, 2007.

Galle, B., Klemedtsson, L., and Griffith, D. W. T.: Application of a Fourier transform IR system for measurements of N2O fluxes using micrometeorological methods, an ultralarge chamber system, and conventional field chambers, *Journal of Geophysical Research-Atmospheres*, 99, 16 575–16 583, https://doi.org/10.1029/94JD00264, 1994.

Gastegger, M., Behler, J., and Marquetand, P.: Machine learning molecular dynamics for the simulation of infrared spectra, *Chemical Science*, 8, 6924–6935, https://doi.org/10.1039/C7SC02267K, 2017.

Gasteiger, J.: The central role of chemoinformatics, *Chemometrics and Intelligent Laboratory Systems*, 82, 200–209, https://doi.org/10.1016/j.chemolab.2005.06.022, 2006.

Ge, Z. and Song, Z.: Nonlinear Probabilistic Monitoring Based on the Gaussian Process Latent Variable Model, *Industrial & Engineering Chemistry Research*, 49, 4792–4799, https://doi.org/10.1021/ie9019402, 2010.

Geisser, S.: The Predictive Sample Reuse Method with Applications, *Journal of the American Statistical Association*, 70, 320–328, https://doi.org/10.1080/01621459.1975.10479865, 1975.

Geladi, P. and Kowalski, B. R.: Partial least-squares regression: a tutorial, *Analytica Chimica Acta*, 185, 1–17, https://doi.org/10.1016/0003-2670(86)80028-9, 1986.

Gibson, E. R., Hudson, P. K., and Grassian, V. H.: Physicochemical properties of nitrate aerosols: Implications for the atmosphere, *Journal of Physical Chemistry A*, 110, 11 785–11 799, https://doi.org/10.1021/jp063821k, 2006.

Gilardoni, S., Russell, L. M., Sorooshian, A., Flagan, R. C., Seinfeld, J. H., Bates, T. S., Quinn, P. K., Allan, J. D., Williams, B., Goldstein, A. H., Onasch, T. B., and Worsnop, D. R.: Regional variation of organic functional groups in aerosol particles on four US east coast platforms during the International Consortium for Atmospheric Research on Transport and Transformation 2004 campaign, *Journal of Geophysical Research-Atmospheres*, 112, D10S27, https://doi.org/10.1029/2006JD007737, 2007.

Gosselin, R., Rodrigue, D., and Duchesne, C.: A Bootstrap-VIP approach for selecting wavelength intervals in spectral imaging applications, *Chemometrics and Intelligent Laboratory Systems*, 100, 12–21, 2010.

Gowen, A. A., Downey, G., Esquerre, C., and O'Donnell, C. P.: Preventing over-fitting in PLS calibration models of near-infrared (NIR) spectroscopy data using regression coefficients, *Journal of Chemometrics*, 25, 375–381, https://doi.org/10.1002/cem.1349, 2011.

Gribov, L. A. and Elyashberg, M. E.: Symbolic logic methods for spectrochemical investigations, *Journal of Molecular Structure*, 5, 179–198, https://doi.org/10.1016/0022-2860(70)80002-3, 1970.

Griffith, D. W. T.: Synthetic Calibration and Quantitative Analysis of Gas-Phase FT-IR Spectra, *Applied Spectroscopy*, 50, 59–70, https://doi.org/10.1366/0003702963906627, 1996.

Griffith, D. W. T. and Galle, B.: Flux measurements of NH3, N2O and CO2 using dual beam FTIR spectroscopy and the flux–gradient technique, *Atmospheric Environment*, 34, 1087–1098, https://doi.org/10.1016/S1352-2310(99)00368-4, 2000.

Griffith, D. W. T. and Jamie, I. M.: Fourier Transform Infrared Spectrometry in Atmospheric and Trace Gas Analysis, in: Encyclopedia of Analytical Chemistry, John Wiley & Sons, Ltd, https://doi.org/10.1002/9780470027318.a0710/abstract, 2006.

Griffith, D. W. T., Leuning, R., Denmead, O. T., and Jamie, I. M.: Air–land exchanges of CO2, CH4 and N2O measured by FTIR spectrometry and micrometeorological techniques, *Atmospheric Environment*, 36, 1833–1842, https://doi.org/10.1016/S1352-2310(02)00139-5, 2002.

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

Griffiths, P. R.: Introduction to Vibrational Spectroscopy, John Wiley & Sons, Ltd, https://doi.org/10.1002/0470027320.s0102, 2006.

Gujral, P., Amrhein, M., Ergon, R., Wise, B. M., and Bonvin, D.: On multivariate calibration with unlabeled data, *Journal of Chemometrics*, 25, 456–465, https://doi.org/10.1002/cem.1389, 2011.

Gussoni, M., Castiglioni, C., and Zerbi, G.: Vibrational Intensities: Interpretation and Use for Diagnostic Purposes, in: Handbook of Vibrational Spectroscopy, John Wiley & Sons, Ltd, https://doi.org/10.1002/0470027320.s4205, 2006.

Halevy, A., Norvig, P., and Pereira, F.: The Unreasonable Effectiveness of Data, *IEEE Intelligent Systems*, 24, 8–12, https://doi.org/10.1109/MIS.2009.36, 2009.

Hallquist, M., Wenger, J. C., Baltensperger, U., Rudich, Y., Simpson, D., Claeys, M., Dommen, J., Donahue, N. M., George, C., Goldstein, A. H., Hamilton, J. F., Herrmann, H., Hoffmann, T., Iinuma, Y., Jang, M., Jenkin, M. E., Jimenez, J. L., Kiendler-Scharr, A., Maenhaut, W., McFiggans, G., Mentel, T. F., Monod, A., Prevot, A. S. H., Seinfeld, J. H., Surratt, J. D., Szmigielski, R., and Wildt, J.: The formation, properties and impact of secondary organic aerosol: current and emerging issues, *Atmospheric Chemistry and Physics*, 9, 5155–5236, https://doi.org/10.5194/acp-9-5155-2009, 2009.

Hammer, S., Griffith, D. W. T., Konrad, G., Vardag, S., Caldow, C., and Levin, I.: Assessment of a multi-species in situ FTIR for precise atmospheric greenhouse gas observations, *Atmospheric Measurement Techniques*, 6, 1153–1170, https://doi.org/10.5194/amt-6-1153-2013, 2013.

Hanst, P. L., Wong, N. W., and Bragin, J.: A long-path infra-red study of Los Angeles smog, *Atmospheric Environment (1967)*, 16, 969–981, https://doi.org/10.1016/0004-6981(82)90183-4, 1982.

Harrington, P. d. B., Urbas, A., and Wan, C.: Evaluation of Neural Network Models with Generalized Sensitivity Analysis, *Analytical Chemistry*, 72, 5004–5013, https://doi.org/10.1021/ac0004963, 2000.

Hase, F., Frey, M., Blumenstock, T., Groß, J., Kiel, M., Kohlhepp, R., Mengistu Tsidu, G., Schäfer, K., Sha, M. K., and Orphal, J.: Application of portable FTIR spectrometers for detecting greenhouse gas emissions of the major city Berlin, *Atmospheric Measurement Techniques*, 8, 3059–3068, https://doi.org/10.5194/amt-8-3059-2015, 2015.

Hasegawa, T.: Principal Component Regression and Partial Least Squares Modeling, in: Handbook of Vibrational Spectroscopy, John Wiley & Sons, Ltd, https://doi.org/10.1002/0470027320.s4604, 2006.

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

Hawkins, L. N. and Russell, L. M.: Oxidation of ketone groups in transported biomass burning aerosol from the 2008 Northern California Lightning Series fires, *Atmospheric Environment*, 44, 4142–4154, https://doi.org/10.1016/j.atmosenv.2010.07.036, 2010.

Hayes, M. H.: Statistical Digital Signal Processing and Modeling, John Wiley & Sons, Inc., New York, NY, USA, 1st edn., 1996.

Hazama, K. and Kano, M.: Covariance-based locally weighted partial least squares for high-performance adaptive modeling, *Chemometrics and Intelligent Laboratory Systems*, 146, 55–62, https://doi.org/10.1016/j.chemolab.2015.05.007, 2015.

Helland, K., Berntsen, H. E., Borgen, O. S., and Martens, H.: Recursive algorithm for partial least squares regression, *Chemometrics and Intelligent Laboratory Systems*, 14, 129–137, https://doi.org/10.1016/0169-7439(92)80098-O, 1992.

Hemmer, M. C.: Expert Systems in Chemistry Research, Taylor & Francis, Inc., Bristol, PA, USA, 2007.

Henry, R. C., Lewis, C. W., Hopke, P. K., and Williamson, H. J.: Review of receptor model fundamentals, *Atmospheric Environment (1967)*, 18, 1507–1515, https://doi.org/10.1016/0004-6981(84)90375-5, 1984.

Hoerl, A. E. and Kennard, R. W.: Ridge Regression - Applications To Nonorthogonal Problems, *Technometrics*, 12, 69–&, https://doi.org/10.2307/1267352, 1970.

Holes, A., Eusebi, A., Grosjean, D., and Allen, D. T.: FTIR analysis of aerosol formed in the photooxidation of 1,3,5-trimethylbenzene, *Aerosol Science and Technology*, 26, 516–526, https://doi.org/10.1080/02786829708965450, 1997.

Hopke, P. K.: Target transformation factor analysis, *Chemometrics and Intelligent Laboratory Systems*, 6, 7–19, https://doi.org/10.1016/0169-7439(89)80061-9, 1989.

Höskuldsson, A.: Prediction Methods in Science and Technology: Basic Theory, vol. 1, Thor Publishing, 1996.

Höskuldsson, A.: Variable and subset selection in PLS regression, *Chemometrics and Intelligent Laboratory Systems*, 55, 23–38, https://doi.org/10.1016/S0169-7439(00)00113-1, 2001.

Hotelling, H.: The Generalization of Student's Ratio, *The Annals of Mathematical Statistics*, 2, 360–378, https://doi.org/10.1214/aoms/1177732979, 1931.

Huang, J., Smola, A. J., Gretton, A., Borgwardt, K. M., and Scholkopf, B.: Correcting Sample Selection Bias by Unlabeled Data, in: Proceedings of the 19th International Conference on Neural Information Processing Systems, NIPS'06, pp. 601–608, MIT Press, Cambridge, MA, USA, http://dl.acm.org/citation.cfm?id=2976456.2976532, 2006.

Huber, P. J. and Ronchetti, E. M.: Robust Statistics, Wiley Series in Probability and Statistics, John Wiley & Sons, Ltd., 2nd ed. edn., https://doi.org/10.1002/9780470434697, 2009.

Hung, H.-M., Chen, Y.-Q., and Martin, S. T.: Reactive Aging of Films of Secondary Organic Material Studied by Infrared Spectroscopy, *The Journal of Physical Chemistry A*, 117, 108–116, https://doi.org/10.1021/jp309470z, 2013.

Hurst, D. F., Griffith, D. W. T., and Cook, G. D.: Trace gas emissions from biomass burning in tropical Australian savannas, *Journal of Geophysical Research-Atmospheres*, 99, 16 441–16 456, https://doi.org/10.1029/94JD00670, 1994.

Isaksson, T. and Aastveit, A. H.: Classification Methods, in: Handbook of Vibrational Spectroscopy, John Wiley & Sons, Ltd, https://doi.org/10.1002/0470027320.s4304, 2006.

Ishiyama, T. and Morita, A.: Molecular Dynamics Simulation of Sum Frequency Generation Spectra of Aqueous Sulfuric Acid Solution, *Journal of Physical Chemistry C*, 115, 13 704–13 716, https://doi.org/10.1021/jp200269k, 2011.

Ivanov, S. D., Witt, A., and Marx, D.: Theoretical spectroscopy using molecular dynamics: theory and application to CH5+ and its isotopologues, *Physical Chemistry Chemical Physics*, 15, 10 270–10 299, https://doi.org/10.1039/C3CP44523B, 2013.

Jackson, J. E.: A User's Guide to Principal Components, Wiley Series in Probability and Statistics, John Wiley & Sons, https://doi.org/10.1002/0471725331, 2004.

Janson, L., Fithian, W., and Hastie, T. J.: Effective degrees of freedom: a flawed metaphor, *Biometrika*, 102, 479–485, https://doi.org/10.1093/biomet/asv019, 2015.

Johnson, N. L.: Systems of Frequency Curves Generated By Methods of Translation, *Biometrika*, 36, 149–176, 1949.

Jouan-Rimbaud, D., Bouveresse, E., Massart, D. L., and de Noord, O. E.: Detection of prediction outliers and inliers in multivariate calibration, *Analytica Chimica Acta*, 388, 283–301, https://doi.org/10.1016/S0003-2670(98)00626-6, 1999.

Kalivas, J. H.: Overview of two-norm (L2) and one-norm (L1) Tikhonov regularization variants for full wavelength or sparse spectral multivariate calibration models or maintenance, *Journal of Chemometrics*, 26, 218–230, https://doi.org/10.1002/cem.2429, 2012.

Kariya, T. and Kurata, H.: Generalized Least Squares, Wiley Series in Probability and Statistics, Wiley, 2004.

Kelley, A. M.: Condensed-Phase Molecular Spectroscopy and Photophysics, John Wiley & Sons, 2013.

Kennard, R. W. and Stone, L. A.: Computer Aided Design of Experiments, *Technometrics*, 11, 137–148, https://doi.org/10.1080/00401706.1969.10490666, 1969.

Kidd, C., Perraud, V., and Finlayson-Pitts, B. J.: New insights into secondary organic aerosol from the ozonolysis of α-pinene from combined infrared spectroscopy and mass spectrometry measurements, *Physical Chemistry Chemical Physics*, 16, 22 706–22 716, https://doi.org/10.1039/C4CP03405H, 2014.

Kim, J., Shusterman, A. A., Lieschke, K. J., Newman, C., and Cohen, R. C.: The BErkeley Atmospheric CO_2 Observation Network: field calibration and evaluation of low-cost air quality sensors, *Atmospheric Measurement Techniques*, 11, 1937–1946, https://doi.org/10.5194/amt-11-1937-2018, 2018.

Kim, S., Kano, M., Nakagawa, H., and Hasebe, S.: Estimation of active pharmaceutical ingredients content using locally weighted partial least squares and statistical wavelength selection, *International Journal of Pharmaceutics*, 421, 269–274, https://doi.org/10.1016/j.ijpharm.2011.10.007, 2011.

Kirchgessner, D. A., Piccot, S. D., and Chadha, A.: Estimation of methane emissions from a surface coal mine using open-path FTIR spectroscopy and modeling techniques, *Chemosphere*, 26, 23–44, https://doi.org/10.1016/0045-6535(93)90410-7, 1993.

Kirchner, U., Scheer, V., and Vogt, R.: FTIR Spectroscopic Investigation of the Mechanism and Kinetics of the Heterogeneous Reactions of NO2 and HNO3 with Soot, *The Journal of Physical Chemistry A*, 104, 8908–8915, https://doi.org/10.1021/jp0005322, 2000.

Koop, T., Bookhold, J., Shiraiwa, M., and Poeschl, U.: Glass transition and phase state of organic compounds: dependency on molecular properties and implications for secondary organic aerosols in the atmosphere, *Physical Chemistry Chemical Physics*, 13, 19 238–19 255, https://doi.org/10.1039/c1cp22617g, 2011.

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

Kourti, T. and MacGregor, J. F.: Process analysis, monitoring and diagnosis, using multivariate projection methods, *Chemometrics and Intelligent Laboratory Systems*, 28, 3–21, https://doi.org/10.1016/0169-7439(95)80036-9, 1995.

Krämer, N. and Sugiyama, M.: The Degrees of Freedom of Partial Least Squares Regression, *Journal of the American Statistical Association*, 106, 697–705, https://doi.org/10.1198/jasa.2011.tm10107, 2011.

Krost, K. J. and McClenny, W. A.: Fourier Transform Infrared Spectrometric Analysis for Particle-Associated Ammonium Sulfate, *Applied Spectroscopy*, 46, 1737–1740, https://doi.org/10.1366/0003702924926763, 1992.

Krost, K. J. and McClenny, W. A.: FT-IR Transmission Spectroscopy for Quantitation of Ammonium Bisulfate in Fine-Particulate Matter Collected on Teflon Filters, *Applied Spectroscopy*, 48, 702–705, https://doi.org/10.1366/000370294774368983, 1994.

Kubicki, J. D. and Mueller, K. T.: Computational Spectroscopy in Environmental Chemistry, in: Computational Spectroscopy, pp. 323–351, Wiley-VCH Verlag GmbH & Co. KGaA, https://doi.org/10.1002/9783527633272.ch11, 2010.

Kuhn, M. and Johnson, K.: Applied Predictive Modeling, SpringerLink : Bücher, Springer New York, https://doi.org/10.1007/978-1-4614-6849-3, 2013.

Kulkarni, A. D., Rai, D., Bartolotti, L. J., and Pathak, R. K.: Microsolvation of methyl hydrogen peroxide: Ab initio quantum chemical approach, *The Journal of Chemical Physics*, 131, 054 310, https://doi.org/10.1063/1.3179753, 2009.

Kulkarni, P., Baron, P. A., and Willeke, K.: Aerosol Measurement: Principles, Techniques, and Applications, John Wiley & Sons, 2011.

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.

Kvalheim, O. M.: Interpretation of partial least squares regression models by means of target projection and selectivity ratio plots, *Journal of Chemometrics*, 24, 496–504, https://doi.org/10.1002/cem.1289, 2010.

Lack, D. A., Moosmueller, H., McMeeking, G. R., Chakrabarty, R. K., and Baumgardner, D.: Characterizing elemental, equivalent black, and refractory black carbon aerosol particles: a review of techniques, their limitations and uncertainties, *Analytical and Bioanalytical Chemistry*, 406, 99–122, https://doi.org/10.1007/s00216-013-7402-3, 2014.

Laskin, J., Laskin, A., and Nizkorodov, S. A.: Mass Spectrometry Analysis in Atmospheric Chemistry, *Analytical Chemistry*, 90, 166–189, https://doi.org/10.1021/acs.analchem.7b04249, 2018.

Latecki, L. J., Lazarevic, A., and Pokrajac, D.: Outlier Detection with Kernel Density Functions, in: Machine Learning and Data Mining in Pattern Recognition, pp. 61–75, Springer, Berlin, Heidelberg, https://doi.org/10.1007/978-3-540-73499-4_6, 2007.

Leardi, R.: Application of genetic algorithm-PLS for feature selection in spectral data sets, *Journal of Chemometrics*, 14, 643–655, 2000.

Leardi, R. and Nørgaard, L.: Sequential application of backward interval partial least squares and genetic algorithms for the selection of relevant spectral regions, *Journal of Chemometrics*, 18, 486–497, https://doi.org/10.1002/cem.893, 2004.

Lee, E., Chan, C. K., and Paatero, P.: Application of positive matrix factorization in source apportionment of particulate pollutants in Hong Kong, *Atmospheric Environment*, 33, 3201–3212, https://doi.org/10.1016/S1352-2310(99)00113-2, 1999.

Li, B., Morris, J., and Martin, E. B.: Model selection for partial least squares regression, *Chemometrics and Intelligent Laboratory Systems*, 64, 79–89, 2002.

Li, Y.-J., Liu, P.-F., Bergoend, C., Bateman, A. P., and Martin, S. T.: Rebounding hygroscopic inorganic aerosol particles: Liquids, gels, and hydrates, *Aerosol Science and Technology*, 51, 388–396, https://doi.org/10.1080/02786826.2016.1263384, 2017.

Lin, Z., Pei, Y., Chen, Z., Shi, X., Qiao, Y., Shi, X., and Qiao, Y.: Improving the creditability and reproducibility of variables selected from near infrared spectra, in: 2013 Ninth International Conference on Natural Computation (ICNC), pp. 1370–1376, https://doi.org/10.1109/ICNC.2013.6818193, 2013.

Lindgren, F., Geladi, P., and Wold, S.: The Kernel Algorithm For PLS, *Journal of Chemometrics*, 7, 45–59, https://doi.org/10.1002/cem.1180070104, 1993.

Liu, F. T., Ting, K. M., and Zhou, Z. H.: Isolation Forest, in: 2008 Eighth IEEE International Conference on Data Mining, pp. 413–422, https://doi.org/10.1109/ICDM.2008.17, 2008.

Liu, J.: Developing a soft sensor based on sparse partial least squares with variable selection, *Journal of Process Control*, 24, 1046–1056, https://doi.org/10.1016/j.jprocont.2014.05.014, 2014.

Long, J. R., Gregoriou, V. G., and Gemperline, P. J.: Spectroscopic calibration and quantitation using artificial neural networks, *Analytical Chemistry*, 62, 1791–1797, https://doi.org/10.1021/ac00216a013, 1990.

Luinge, H. J., van der Maas, J. H., and Visser, T.: Partial least squares regression as a multivariate tool for the interpretation of infrared spectra, *Chemometrics and Intelligent Laboratory Systems*, 28, 129–138, https://doi.org/10.1016/0169-7439(95)80045-B, 1995.

Ma, Y., Gong, W., and Mao, F.: Transfer learning used to analyze the dynamic evolution of the dust aerosol, *Journal of Quantitative Spectroscopy and Radiative Transfer*, 153, 119–130, https://doi.org/10.1016/j.jqsrt.2014.09.025, 2015.

MacDonald, S. A. and Bureau, B.: Fourier Transform Infrared Attenuated Total Reflection and Transmission Spectra Studied by Dispersion Analysis, *Applied Spectroscopy*, 57, 282–287, 2003.

MacGregor, J. F. and Kourti, T.: Statistical process control of multivariate processes, *Control Engineering Practice*, 3, 403–414, https://doi.org/10.1016/0967-0661(95)00014-L, 1995.

Mader, P. P., MacPhee, R. D., Lofberg, R. T., and Larson, G. P.: Composition of Organic Portion of Atmospheric Aerosols in the Los Angeles Area, *Industrial & Engineering Chemistry*, 44, 1352–1355, https://doi.org/10.1021/ie50510a047, 1952.

Mahalanobis, P.: On the Generalised Distance in Statistics, *In Proceedings National Institute of Science, India*, 2, 49–55, 1936.

Malli, B., Birlutiu, A., and Natschläger, T.: Standard-free calibration transfer - An evaluation of different techniques, *Chemometrics and Intelligent Laboratory Systems*, 161, 49–60, https://doi.org/10.1016/j.chemolab.2016.12.008, 2017.

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., Schichtel, B. A., and Pitchford, M. L.: Uncertainties in PM2.5 Gravimetric and Speciation Measurements and What We Can Learn from Them, *Journal of the Air & Waste Management Association*, 61, 1131–1149, https://doi.org/10.1080/10473289.2011.603998, 2011.

Marcou, G., Delouis, G., Mokshyna, O., Horvath, D., Lachiche, N., and Varnek, A.: Transductive Ridge Regression in Structure-Activity Modeling, *Molecular Informatics*, 36, 1700 112, https://doi.org/10.1002/minf.201700112, 2017.

Maria, S. F., Russell, L. M., Turpin, B. J., and Porcja, R. J.: FTIR measurements of functional groups and organic mass in aerosol samples over the Caribbean, *Atmospheric Environment*, 36, 5185–5196, https://doi.org/10.1016/S1352-2310(02)00654-4, 2002.

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.

Marsalek, O. and Markland, T. E.: Quantum Dynamics and Spectroscopy of Ab Initio Liquid Water: The Interplay of Nuclear and Electronic Quantum Effects, *The Journal of Physical Chemistry Letters*, 8, 1545–1551, https://doi.org/10.1021/acs.jpclett.7b00391, 2017.

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

Marx, D.: Ab Initio Molecular Dynamics: Basic Theory and Advanced Methods, Cambridge University Press, Cambridge, UK; New York, 1st edn., 2009.

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.

Medders, G. R. and Paesani, F.: Infrared and Raman Spectroscopy of Liquid Water through "First-Principles" Many-Body Molecular Dynamics, *Journal of Chemical Theory and Computation*, 11, 1145–1154, https://doi.org/10.1021/ct501131j, 2015.

Mehmood, T., Liland, K. H., Snipen, L., and Saebo, S.: A review of variable selection methods in Partial Least Squares Regression, *Chemometrics and Intelligent Laboratory Systems*, 118, 62–69, https://doi.org/10.1016/j.chemolab.2012.07.010, 2012.

Meier, A. and Notholt, J.: Determination of the isotopic abundances of heavy O3 as observed in Arctic ground-based FTIR-spectra, *Geophysical Research Letters*, 23, 551–554, https://doi.org/10.1029/96GL00374, 1996.

Mevik, B. and Wehrens, R.: The pls package: Principal component and partial least squares regression in R, *Journal of Statistical Software*, 18, 1–24, https://doi.org/10.18637/jss.v018.i02, 2007.

Molinaro, A. M., Simon, R., and Pfeiffer, R. M.: Prediction error estimation: a comparison of resampling methods, *Bioinformatics*, 21, 3301–3307, https://doi.org/10.1093/bioinformatics/bti499, 2005.

Montgomery, D.: Statistical Quality Control, John Wiley & Sons, 7th edition edn., 2013.

Mosteller, F. and Tukey, J.: Data Analysis, including Statistics, in: Revised Handbook of Social Psychology, edited by Lindzey, G. and Aronson, E., vol. 2, pp. 80–203, Addison Wesley, 1968.

Munk, M. E.: Computer-Based Structure Determination: Then and Now, *Journal of Chemical Information and Computer Sciences*, 38, 997–1009, https://doi.org/10.1021/ci980083r, 1998.

Murphy, K. P.: Machine Learning: A Probabilistic Perspective, Adaptive computation and machine learning, MIT Press, 2012.

Mylonas, D. T., Allen, D. T., Ehrman, S. H., and Pratsinis, S. E.: The Sources and Size Distributions of Organonitrates In Los Angeles Aerosol, *Atmospheric Environment Part A-general Topics*, 25, 2855–2861, https://doi.org/10.1016/0960-1686(91)90211-O, 1991.

Nelder, J. A. and Wedderburn, R. W. M.: Generalized Linear Models, *Journal of the Royal Statistical Society. Series A (General)*, 135, 370–384, 1972.

Nomikos, P. and MacGregor, J. F.: Multivariate SPC Charts for Monitoring Batch Processes, *Technometrics*, 37, 41–59, https://doi.org/10.1080/00401706.1995.10485888, 1995.

Nordlund, T. M.: Quantitative Understanding of Biosystems: An Introduction to Biophysics, CRC Press, 2011.

Novakov, T.: The role of soot and primary oxidants in atmospheric chemistry, *Science of The Total Environment*, 36, 1–10, https://doi.org/10.1016/0048-9697(84)90241-9, 1984.

Novic, M. and Zupan, J.: Investigation of Infrared Spectra-Structure Correlation Using Kohonen and Counterpropagation Neural Network, *Journal of Chemical Information and Computer Sciences*, 35, 454–466, https://doi.org/10.1021/ci00025a013, 1995.

Nozière, B., Kalberer, M., Claeys, M., Allan, J., D'Anna, B., Decesari, S., Finessi, E., Glasius, M., Grgić, I., Hamilton, J. F., Hoffmann, T., Iinuma, Y., Jaoui, M., Kahnt, A., Kampf, C. J., Kourtchev, I., Maenhaut, W., Marsden, N., Saarikoski, S., Schnelle-Kreis, J., Surratt, J. D., Szidat, S., Szmigielski, R., and Wisthaler, A.: The Molecular Identification of Organic Compounds in the Atmosphere: State of the Art and Challenges, *Chemical Reviews*, 115, 3919–3983, https://doi.org/10.1021/cr5003485, 2015.

Ofner, J.: Formation of secondary organic aerosol and its processing by atmospheric halogen species — a spectroscopic study, Ph.D. thesis, University of Bayreuth, http://opus.ub.uni-bayreuth.de/volltexte/2011/915/, 2011.

Olivieri, A. C.: Practical guidelines for reporting results in single- and multi-component analytical calibration: A tutorial, *Analytica Chimica Acta*, 868, 10–22, https://doi.org/10.1016/j.aca.2015.01.017, 2015.

Olivieri, A. C., Faber, N. M., Ferré, J., Boqué, R., Kalivas, J. H., and Mark, H.: Uncertainty estimation and figures of merit for multivariate calibration (IUPAC Technical Report), *Pure and Applied Chemistry*, 78, 633–661, https://doi.org/10.1351/pac200678030633, 2006.

Oppenheimer, C. and Kyle, P. R.: Probing the magma plumbing of Erebus volcano, Antarctica, by open-path FTIR spectroscopy of gas emissions, *Journal of Volcanology and Geothermal Research*, 177, 743–754, https://doi.org/10.1016/j.jvolgeores.2007.08.022, 2008.

Ottaway, J., Farrell, J. A., and Kalivas, J. H.: Spectral Multivariate Calibration without Laboratory Prepared or Determined Reference Analyte Values, *Analytical Chemistry*, 85, 1509–1516, https://doi.org/10.1021/ac302705m, 2012.

Paatero, P.: Least squares formulation of robust non-negative factor analysis, *Chemometrics and Intelligent Laboratory Systems*, 37, 23–35, https://doi.org/10.1016/S0169-7439(96)00044-5, 1997.

Pagliai, M., Cavazzoni, C., Cardini, G., Erbacci, G., Parrinello, M., and Schettino, V.: Anharmonic infrared and Raman spectra in Car-Parrinello molecular dynamics simulations, *The Journal of Chemical Physics*, 128, 224 514, https://doi.org/10.1063/1.2936988, 2008.

Painter, P. C., Snyder, R. W., Starsinic, M., Coleman, M. M., Kuehn, D. W., and Davis, A.: Fourier Transform IR Spectroscopy, in: Coal and Coal Products: Analytical Characterization Techniques, vol. 205 of *ACS Symposium Series*, pp. 47–76, AMERICAN CHEMICAL SOCIETY, https://doi.org/10.1021/bk-1982-0205.ch003, dOI: 10.1021/bk-1982-0205.ch003, 1982.

Paiva, J. G. S., Schwartz, W. R., Pedrini, H., and Minghim, R.: Semi-Supervised Dimensionality Reduction based on Partial Least Squares for Visual Analysis of High Dimensional Data, *Computer Graphics Forum*, 31, 1345–1354, https://doi.org/10.1111/j.1467-8659.2012.03126.x, 2012.

Palen, E. J., Allen, D. T., Pandis, S. N., Paulson, S. E., Seinfeld, J. H., and Flagan, R. C.: Fourier-transform Infrared-analysis of Aerosol Formed In the Photooxidation of Isoprene and Beta-pinene, *Atmospheric Environment Part A-general Topics*, 26, 1239–1251, https://doi.org/10.1016/0960-1686(92)90385-X, 1992.

Palen, E. J., Allen, D. T., Pandis, S. N., Paulson, S., Seinfeld, J. H., and Flagan, R. C.: Fourier-transform Infrared-analysis of Aerosol Formed In the Photooxidation of 1-octene, *Atmospheric Environment Part A-General Topics*, 27, 1471–1477, https://doi.org/10.1016/0960-1686(93)90133-J, 1993.

Pan, S. J. and Yang, Q.: A Survey on Transfer Learning, *IEEE Transactions on Knowledge and Data Engineering*, 22, 1345–1359, https://doi.org/10.1109/TKDE.2009.191, 2010.

Pan, S. J., Tsang, I. W., Kwok, J. T., and Yang, Q.: Domain Adaptation via Transfer Component Analysis, *IEEE Transactions on Neural Networks*, 22, 199–210, https://doi.org/10.1109/TNN.2010.2091281, 2011.

Paulson, S. E., Pandis, S. N., Baltensperger, U., Seinfeld, J. H., Flagan, R. C., Palen, E. J., Allen, D. T., Schaffner, C., Giger, W., and Portmann, A.: Characterization of Photochemical Aerosols From Biogenic Hydrocarbons, *Journal of Aerosol Science*, 21, GESELL AEROSOL-FORSCH; ETH, https://doi.org/10.1016/0021-8502(90)90230-U, 1990.

Pedone, A., Biczysko, M., and Barone, V.: Environmental Effects in Computational Spectroscopy: Accuracy and Interpretation, *ChemPhysChem*, 11, 1812–1832, https://doi.org/10.1002/cphc.200900976, 2010.

Petzold, A., Ogren, J. A., Fiebig, M., Laj, P., Li, S. . M., Baltensperger, U., Holzer-Popp, T., Kinne, S., Pappalardo, G., Sugimoto, N., Wehrli, C., Wiedensohler, A., and Zhang, X. . Y.: Recommendations for reporting "black carbon" measurements, *Atmospheric Chemistry and Physics*, 13, 8365–8379, https://doi.org/10.5194/acp-13-8365-2013, 2013.

Phatak, A., Reilly, P. M., and Penlidis, A.: An approach to interval estimation in partial least squares regression, *Analytica Chimica Acta*, 277, 495–501, https://doi.org/10.1016/0003-2670(93)80461-S, 1993.

Pickle, T., Allen, D. T., and Pratsinis, S. E.: The sources and size distributions of aliphatic and carbonyl carbon in Los Angeles aerosol, *Atmospheric Environment. Part A. General Topics*, 24, 2221–2228, https://doi.org/10.1016/0960-1686(90)90253-J, 1990.

Pimentel, M. A., Clifton, D. A., Clifton, L., and Tarassenko, L.: A review of novelty detection, *Signal Processing*, 99, 215–249, https://doi.org/10.1016/j.sigpro.2013.12.026, 2014.

Pitts, J. N., Finlayson-Pitts, B. J., and Winer, A. M.: Optical systems unravel smog chemistry, *Environmental Science & Technology*, 11, 568–573, https://doi.org/10.1021/es60129a014, 1977.

Pitts, J. N., Sanhueza, E., Atkinson, R., Carter, W. P. L., Winer, A. M., Harris, G. W., and Plum, C. N.: An investigation of the dark formation of nitrous acid in environmental chambers, *International Journal of Chemical Kinetics*, 16, 919–939, https://doi.org/10.1002/kin.550160712, 1984.

Pollard, M., Jaklevic, J., and Howes, J.: Fourier Transform Infrared and Ion-Chromatographic Sulfate Analysis of Ambient Air Samples, *Aerosol Science and Technology*, 12, 105–113, https://doi.org/10.1080/02786829008959330, 1990.

Popovicheva, O. B., Kireeva, E. D., Shonija, N. K., Vojtisek-Lom, M., and Schwarz, J.: FTIR analysis of surface functionalities on particulate matter produced by off-road diesel engines operating on diesel and biofuel, *Environmental Science and Pollution Research*, 22, 4534–4544, https://doi.org/10.1007/s11356-014-3688-8, 2014.

Pratt, K. A. and Prather, K. A.: Mass spectrometry of atmospheric aerosolsuRecent developments and applications. Part I: Off-line mass spectrometry techniques, *Mass Spectrometry Reviews*, 31, 1–16, https://doi.org/10.1002/mas.20322, 2012.

Presto, A. A., Hartz, K. E. H., and Donahue, N. M.: Secondary organic aerosol production from terpene ozonolysis. 2. Effect of NOx concentration, *Environmental Science & Technology*, 39, 7046–7054, https://doi.org/10.1021/es050400s, 2005.

Putrino, A. and Parrinello, M.: Anharmonic Raman Spectra in High-Pressure Ice from Ab Initio Simulations, *Physical Review Letters*, 88, 176 401, https://doi.org/10.1103/PhysRevLett.88.176401, 2002.

Qin, S. J.: Recursive PLS algorithms for adaptive data modeling, *Computers & Chemical Engineering*, 22, 503–514, https://doi.org/10.1016/S0098-1354(97)00262-7, 1998.

Quarti, C., Milani, A., and Castiglioni, C.: Ab Initio Calculation of the IR Spectrum of PTFE: Helical Symmetry and Defects, *The Journal of Physical Chemistry B*, 117, 706–718, https://doi.org/10.1021/jp3102145, 2013.

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.

Rinnan, Å.: Pre-processing in vibrational spectroscopy — when, why and how, *Analytical Methods*, 6, 7124–7129, https://doi.org/10.1039/C3AY42270D, 2014.

Rinnan, Å., Nørgaard, L., Berg, F. v. d., Thygesen, J., Bro, R., and Engelsen, S. B.: Chapter 2 - Data Pre-processing, in: Infrared Spectroscopy for Food Quality Analysis and Control, edited by Sun, D.-W., pp. 29–50, Academic Press, San Diego, http://www.sciencedirect.com/science/article/pii/B978012374136300002X, 2009.

Robb, E. W. and Munk, M. E.: A neural network approach to infrared spectrum interpretation, *Microchimica Acta*, 100, 131–155, https://doi.org/10.1007/BF01244838, 1990.

Rosipal, R. and Krämer, N.: Overview and Recent Advances in Partial Least Squares, in: Subspace, Latent Structure and Feature Selection, edited by Saunders, C., Grobelnik, M., Gunn, S., and Shawe-Taylor, J., vol. 3940 of *Lecture Notes in Computer Science*, pp. 34–51, Springer Berlin Heidelberg, https://doi.org/10.1007/11752790_2, 2006.

Rossi, M., Ceriotti, M., and Manolopoulos, D. E.: How to remove the spurious resonances from ring polymer molecular dynamics, *The Journal of Chemical Physics*, 140, 234 116, https://doi.org/10.1063/1.4883861, 2014a.

Rossi, M., Liu, H., Paesani, F., Bowman, J., and Ceriotti, M.: Communication: On the consistency of approximate quantum dynamics simulation methods for vibrational spectra in the condensed phase, *The Journal of Chemical Physics*, 141, 181 101, https://doi.org/10.1063/1.4901214, 2014b.

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.

Russolillo, G.: Non-Metric Partial Least Squares, *Electronic Journal of Statistics*, 6, 1641–1669, https://doi.org/10.1214/12-EJS724, 2012.

Russwurm, G. M.: Compendium Method TO-16: Long-path Open-path Fourier Transform Infrared Monitoring of Atmospheric Gases, pp. 16.1–16.41, US Environmental Protection Agency, 1999.

Russwurm, G. M. and Childers, J. W.: Open-Path Fourier Transform Infrared Spectroscopy, in: Handbook of Vibrational Spectroscopy, John Wiley & Sons, Ltd, https://doi.org/10.1002/0470027320.s2112, 2006.

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.

Sadezky, A., Muckenhuber, H., Grothe, H., Niessner, R., and Pöschl, U.: Raman microspectroscopy of soot and related carbonaceous materials: Spectral analysis and structural information, *Carbon*, 43, 1731–1742, https://doi.org/10.1016/j.carbon.2005.02.018, 2005.

Saeys, W., De Ketelaere, B., and Darius, P.: Potential applications of functional data analysis in chemometrics, *Journal of Chemometrics*, 22, 335–344, https://doi.org/10.1002/cem.1129, 2008.

Saeys, Y., Inza, I., and Larrañaga, P.: A review of feature selection techniques in bioinformatics, *Bioinformatics*, 23, 2507–2517, https://doi.org/10.1093/bioinformatics/btm344, 2007.

Sasaki, S., Abe, H., Ouki, T., Sakamoto, M., and Ochiai, S.: Automated structure elucidation of several kinds of aliphatic and alicyclic compounds, *Analytical Chemistry*, 40, 2220–2223, https://doi.org/10.1021/ac50158a061, 1968.

Savitzky, A. and Golay, M. J. E.: Smoothing and Differentiation of Data by Simplified Least Squares Procedures., *Analytical Chemistry*, 36, 1627–1639, https://doi.org/10.1021/ac60214a047, 1964.

Sax, M., Zenobi, R., Baltensperger, U., and Kalberer, M.: Time resolved infrared spectroscopic analysis of aerosol formed by photo-oxidation of 1,3,5-trimethylbenzene and alpha-pinene, *Aerosol Science and Technology*, 39, 822–830, https://doi.org/10.1080/02786820500257859, 2005.

Schölkopf, B., Williamson, R., Smola, A., Shawe-Taylor, J., and Platt, J.: Support Vector Method for Novelty Detection, in: Proceedings of the 12th International Conference on Neural Information Processing Systems, NIPS'99, pp. 582–588, MIT Press, Cambridge, MA, USA, http://dl.acm.org/citation.cfm?id=3009657.3009740, 1999.

Schütze, C., Lau, S., Reiche, N., Sauer, U., Borsdorf, H., and Dietrich, P.: Ground-based Remote Sensing with Open-path Fourier-transform Infrared (OP-FTIR) Spectroscopy for Large-scale Monitoring of Greenhouse Gases, *Energy Procedia*, 37, 4276–4282, https://doi.org/10.1016/j.egypro.2013.06.330, 2013.

Schuur, J. and Gasteiger, J.: Infrared Spectra Simulation of Substituted Benzene Derivatives on the Basis of a 3D Structure Representation, *Analytical Chemistry*, 69, 2398–2405, https://doi.org/10.1021/ac9611071, 1997.

Schwarz, G.: Estimating the Dimension of a Model, *The Annals of Statistics*, 6, 461–464, 1978.

Seinfeld, J. and Pandis, S.: Atmospheric Chemistry and Physics: From Air Pollution to Climate Change, John Wiley & Sons, New York, 3rd edn., 2016.

Selzer, P., Gasteiger, J., Thomas, H., and Salzer, R.: Rapid Access to Infrared Reference Spectra of Arbitrary Organic Compounds: Scope and Limitations of an Approach to the Simulation of Infrared Spectra by Neural Networks, *Chemistry – A European Journal*, 6, 920–927, https://doi.org/10.1002/(SICI)1521-3765(20000303)6:5<920::AID-CHEM920>3.0.CO;2-W, 2000.

Serneels, S., Croux, C., and Van Espen, P. J.: Influence properties of partial least squares regression, *Chemometrics and Intelligent Laboratory Systems*, 71, 13–20, https://doi.org/10.1016/j.chemolab.2003.10.009, 2004.

Serradilla, J., Shi, J., and Morris, A.: Fault detection based on Gaussian process latent variable models, *Chemometrics and Intelligent Laboratory Systems*, 109, 9–21, https://doi.org/10.1016/j.chemolab.2011.07.003, 2011.

Shao, L. and Griffiths, P. R.: Information Extraction from a Complex Multicomponent System by Target Factor Analysis, *Analytical Chemistry*, 82, 106–114, https://doi.org/10.1021/ac901246x, 2010.

Shurvell, H.: Spectra–Structure Correlations in the Mid- and Far-Infrared, in: Handbook of Vibrational Spectroscopy, John Wiley & Sons, Ltd, https://doi.org/10.1002/0470027320.s4101, 2006.

Silvestrelli, P. L., Bernasconi, M., and Parrinello, M.: Ab initio infrared spectrum of liquid water, *Chemical Physics Letters*, 277, 478–482, https://doi.org/10.1016/S0009-2614(97)00930-5, 1997.

Solomon, P. A., Crumpler, D., Flanagan, J. B., Jayanty, R., Rickman, E. E., and McDade, C. E.: U.S. National PM2.5 Chemical Speciation Monitoring Networks—CSN and IMPROVE: Description of networks, *Journal of the Air & Waste Management Association*, 64, 1410–1438, https://doi.org/10.1080/10962247.2014.956904, 2014.

Spellicy, R. L. and Webb, J. D.: Atmospheric Monitoring Using Extractive Techniques, in: Handbook of Vibrational Spectroscopy, John Wiley & Sons, Ltd, https://doi.org/10.1002/0470027320.s2111, 2006.

Steele, D.: Infrared Spectroscopy: Theory, in: Handbook of Vibrational Spectroscopy, John Wiley & Sons, Ltd, https://doi.org/10.1002/0470027320.s0103, 2006.

Stone, M.: Cross-Validatory Choice and Assessment of Statistical Predictions, *Journal of the Royal Statistical Society. Series B (Methodological)*, 36, 111–147, 1974.

Sugiyama, M., Nakajima, S., Kashima, H., Buenau, P. V., and Kawanabe, M.: Direct Importance Estimation with Model Selection and Its Application to Covariate Shift Adaptation, in: Advances in Neural Information Processing Systems 20, edited by Platt, J. C., Koller, D., Singer, Y., and Roweis, S. T., pp. 1433–1440, Curran Associates, Inc., 2008.

Takahama, S. and Dillner, A. M.: Model selection for partial least squares calibration and implications for analysis of atmospheric organic aerosol samples with mid-infrared spectroscopy, *Journal of Chemometrics*, 29, 659–668, https://doi.org/10.1002/cem.2761, 2015.

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., 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., 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, 2013.

Takahama, S., Ruggeri, G., and Dillner, A. M.: Analysis of functional groups in atmospheric aerosols by infrared spectroscopy: sparse methods for statistical selection of relevant absorption bands, *Atmospheric Measurement Techniques*, 9, 3429–3454, https://doi.org/10.5194/amt-9-3429-2016, 2016.

Thissen, U., Pepers, M., Üstün, B., Melssen, W. J., and Buydens, L. M. C.: Comparing support vector machines to PLS for spectral regression applications, *Chemometrics and Intelligent Laboratory Systems*, 73, 169–179, 2004.

Thomas, M., Brehm, M., Fligg, R., Vöhringer, P., and Kirchner, B.: Computing vibrational spectra from ab initio molecular dynamics, *Physical Chemistry Chemical Physics*, 15, 6608, https://doi.org/10.1039/c3cp44302g, 2013.

Tibshirani, R.: Regression shrinkage and selection via the Lasso, *Journal of the Royal Statistical Society Series B (Methodological)*, 58, 267–288, 1996.

Tibshirani, R. J.: Degrees of Freedom and Model Search, *ArXiv e-prints*, 2014.

Tikhonov, A. N. and Arsenin, V. I.: Solutions of ill-posed problems, Halsted Press, New York, 1977.

Torrey, L. and Shavlik, J.: Transfer learning, *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*, 1, 242, 2009.

Trygg, J.: O2-PLS for qualitative and quantitative analysis in multivariate calibration, *Journal of Chemometrics*, 16, 283–293, https://doi.org/10.1002/cem.724, 2002.

Tsai, A. C., Liou, M., Simak, M., and Cheng, P. E.: On hyperbolic transformations to normality, *Computational Statistics & Data Analysis*, 115, 250–266, https://doi.org/10.1016/j.csda.2017.06.001, 2017.

Tsai, Y. I. and Kuo, S.-C.: Development of diffuse reflectance infrared Fourier transform spectroscopy for the rapid characterization of aerosols, *Atmospheric Environment*, 40, 1781–1793, https://doi.org/10.1016/j.atmosenv.2005.11.023, 2006.

Tuazon, E. C., Winer, A. M., and Pitts, J. N.: Trace pollutant concentrations in a multiday smog episode in the California South Coast Air Basin by long path length Fourier transform infrared spectroscopy, *Environmental Science & Technology*, 15, 1232–1237, https://doi.org/10.1021/es00092a014, 1981.

Tuinstra, F. and Koenig, J. L.: Raman Spectrum of Graphite, *The Journal of Chemical Physics*, 53, 1126–1130, https://doi.org/10.1063/1.1674108, 1970.

Turrell, G.: Theory of Infrared Spectroscopy, in: Encyclopedia of Analytical Chemistry, John Wiley & Sons, Ltd, https://doi.org/10.1002/9780470027318.a5607/abstract, dOI: 10.1002/9780470027318.a5607, 2006.

U.S. EPA: Method 320 Measurement of vapor phase organic and inorganic emissions by extractive Fourier transform infrared (FTIR) spectroscopy, 1998.

van der Voet, H.: Comparing the predictive accuracy of models using a simple randomization test, *Chemometrics and Intelligent Laboratory Systems*, 25, 313–323, https://doi.org/10.1016/0169-7439(94)85050-X, 1994.

Venables, W. N. and Ripley, B. D.: Modern Applied Statistics with S, Springer, 2003.

Virtanen, A., Joutsensaari, J., Koop, T., Kannosto, J., Yli-Pirila, P., Leskinen, J., Makela, J. M., Holopainen, J. K., Poeschl, U., Kulmala, M., Worsnop, D. R., and Laaksonen, A.: An amorphous solid state of biogenic secondary organic aerosol particles, *Nature*, 467, 824–827, https://doi.org/10.1038/nature09455, 2010.

Walczak, B. and Massart, D.: Local modelling with radial basis function networks, *Chemometrics and Intelligent Laboratory Systems*, 50, 179–198, https://doi.org/10.1016/S0169-7439(99)00056-8, 2000.

Walczak, B. and Wegscheider, W.: Non-linear modelling of chemical data by combinations of linear and neural net methods, *Analytica Chimica Acta*, 283, 508–517, https://doi.org/10.1016/0003-2670(93)85264-K, 1993.

Wang, L.-L., Lin, Y.-W., Wang, X.-F., Xiao, N., Xu, Y.-D., Li, H.-D., and Xu, Q.-S.: A selective review and comparison for interval variable selection in spectroscopic modeling, *Chemometrics and Intelligent Laboratory Systems*, https://doi.org/10.1016/j.chemolab.2017.11.008, 2017.

Weakley, A., Miller, A., Griffiths, P., and Bayman, S.: Quantifying silica in filter-deposited mine dusts using infrared spectra and partial least squares regression, *Analytical and Bioanalytical Chemistry*, 406, 4715–4724, https://doi.org/10.1007/s00216-014-7856-y, 2014.

Weakley, A. T., Takahama, S., and Dillner, A. M.: Ambient aerosol composition by infrared spectroscopy and partial least-squares in the chemical speciation network: Organic carbon with functional group identification, *Aerosol Science and Technology*, 50, 1096–1114, https://doi.org/10.1080/02786826.2016.1217389, 2016.

Weakley, A. T., Takahama, S., and Dillner, A. M.: Thermal/optical reflectance equivalent organic and elemental carbon determined from federal reference and equivalent method fine particulate matter samples using Fourier transform infrared spectrometry, *Aerosol Science and Technology*, 52, 1048–1058, https://doi.org/10.1080/02786826.2018.1504161, 2018a.

Weakley, A. T., Takahama, S., Wexler, A. S., and Dillner, A. M.: Ambient aerosol composition by infrared spectroscopy and partial least squares in the chemical speciation network: Multilevel modeling for elemental carbon, *Aerosol Science and Technology*, 52, 642–654, https://doi.org/10.1080/02786826.2018.1439571, 2018b.

Wei, S., Kulkarni, P., Ashley, K., and Zheng, L.: Measurement of Crystalline Silica Aerosol Using Quantum Cascade Laser-Based Infrared Spectroscopy, *Scientific Reports*, 7, 13 860, https://doi.org/10.1038/s41598-017-14363-3, 2017.

Weigel, U. M. and Herges, R.: Simulation of infrared spectra using artificial neural networks based on semiempirical and empirical data, *Analytica Chimica Acta*, 331, 63–74, https://doi.org/10.1016/0003-2670(96)00203-6, 1996.

Weymuth, T., Haag, M. P., Kiewisch, K., Luber, S., Schenk, S., Jacob, C. R., Herrmann, C., Neugebauer, J., and Reiher, M.: MOVIPAC: Vibrational spectroscopy with a robust meta-program for massively parallel standard and inverse calculations, *Journal of Computational Chemistry*, 33, 2186–2198, https://doi.org/10.1002/jcc.23036, 2012.

Wiklund, S., Nilsson, D., Eriksson, L., Sjostrom, M., Wold, S., and Faber, K.: A randomization test for PLS component selection, *Journal of Chemometrics*, 21, 427–439, https://doi.org/10.1002/cem.1086, 2007.

Wise, B. M. and Gallagher, N. B.: The process chemometrics approach to process monitoring and fault detection, *Journal of Process Control*, 6, 329–348, https://doi.org/10.1016/0959-1524(96)00009-1, 1996.

Wise, B. M. and Roginski, R. T.: A Calibration Model Maintenance Roadmap, *IFAC-PapersOnLine*, 48, 260–265, https://doi.org/10.1016/j.ifacol.2015.08.191, 2015.

Witt, A., Ivanov, S. D., Shiga, M., Forbert, H., and Marx, D.: On the applicability of centroid and ring polymer path integral molecular dynamics for vibrational spectroscopy, *The Journal of Chemical Physics*, 130, 194 510, https://doi.org/10.1063/1.3125009, 2009.

Wold, H.: Estimation of Principal Components and Related Models by Iterative Least squares, in: Multivariate Analysis, pp. 391–420, Academic Press, 1966.

Wold, S.: Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models, *Technometrics*, 20, 397–405, https://doi.org/10.1080/00401706.1978.10489693, 1978.

Wold, S.: Discussion: PLS in Chemical Practice, *Technometrics*, 35, 136–139, https://doi.org/10.2307/1269657, 1993.

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.

Wold, S., Ruhe, A., Wold, H., and Dunn, III, W. J.: The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses, *SIAM Journal on Scientific and Statistical Computing*, 5, 735–743, https://doi.org/10.1137/0905052, 1984.

Wold, S., Antti, H., Lindgren, F., and Öhman, J.: Orthogonal signal correction of near-infrared spectra, *Chemometrics and Intelligent Laboratory Systems*, 44, 175–185, https://doi.org/10.1016/S0169-7439(98)00109-9, 1998.

Wold, S., Trygg, J., Berglund, A., and Antti, H.: Some recent developments in PLS modeling, *Chemometrics and Intelligent Laboratory Systems*, 58, 131–150, https://doi.org/10.1016/S0169-7439(01)00156-3, 2001.

Yao, J., Fan, B., Doucet, J.-P., Panaye, A., Yuan, S., and Li, J.: SIRS-SS: A System for Simulating IR/Raman Spectra. 1. Substructure/Subspectrum Correlation, *Journal of Chemical Information and Computer Sciences*, 41, 1046–1052, https://doi.org/10.1021/ci010010z, 2001.

Yokelson, R. J., Susott, R., Ward, D. E., Reardon, J., and Griffith, D. W. T.: Emissions from smoldering combustion of biomass measured by open-path Fourier transform infrared spectroscopy, *Journal of Geophysical Research-Atmospheres*, 102, 18 865–18 877, https://doi.org/10.1029/97JD00852, 1997.

Zadrozny, B.: Learning and Evaluating Classifiers Under Sample Selection Bias, in: Proceedings of the Twenty-first International Conference on Machine Learning, ICML '04, pp. 114–, ACM, New York, NY, USA, https://doi.org/10.1145/1015330.1015425, 2004.

Zeng, G., Holladay, S., Langlois, D., Zhang, Y., and Liu, Y.: Kinetics of Heterogeneous Reaction of Ozone with Linoleic Acid and its Dependence on Temperature, Physical State, RH, and Ozone Concentration, *The Journal of Physical Chemistry A*, 117, 1963–1974, https://doi.org/10.1021/jp308304n, 2013.

Zezula, P., Amato, G., Dohnal, V., and Batko, M.: Similarity Search: The Metric Space Approach, Advances in Database Systems, Springer US, 2006.

Zhang, L. and Garcia-Munoz, S.: A comparison of different methods to estimate prediction uncertainty using Partial Least Squares (PLS): A practitioner's perspective, *Chemometrics and Intelligent Laboratory Systems*, 97, 152–158, https://doi.org/10.1016/j.chemolab.2009.03.007, 2009.

Zhang, X., Kano, M., and Li, Y.: Locally weighted kernel partial least squares regression based on sparse nonlinear features for virtual sensing of nonlinear time-varying processes, *Computers & Chemical Engineering*, 104, 164–171, https://doi.org/10.1016/j.compchemeng.2017.04.014, 2017.

Zhao, N., Wu, Z.-s., Zhang, Q., Shi, X.-y., Ma, Q., and Qiao, Y.-j.: Optimization of Parameter Selection for Partial Least Squares Model Development, *Scientific Reports*, 5, 11 647, https://doi.org/10.1038/srep11647, 2015.

Zhao, R., Lee, A. K. Y., and Abbatt, J. P. D.: Investigation of Aqueous-Phase Photooxidation of Glyoxal and Methylglyoxal by Aerosol Chemical Ionization Mass Spectrometry: Observation of Hydroxyhydroperoxide Formation, *Journal of Physical Chemistry A*, 116, 6253–6263, https://doi.org/10.1021/jp211528d, 2012.

Zhou, L. M., Hopke, P. K., Stanier, C. O., Pandis, S. N., Ondov, J. M., and Pancras, J. P.: Investigation of the relationship between chemical composition and size distribution of airborne particles by partial least squares and positive matrix factorization, *Journal of Geophysical Research-Atmospheres*, 110, D07S18, https://doi.org/10.1029/2004JD005050, 2005.

Zimmerman, N., Presto, A. A., Kumar, S. P. N., Gu, J., Hauryliuk, A., Robinson, E. S., Robinson, A. L., and Subramanian, R.: A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring, *Atmospheric Measurement Techniques*, 11, 291–313, https://doi.org/10.5194/amt-11-291-2018, 2018.

Ţucureanu, V., Matei, A., and Avram, A. M.: FTIR Spectroscopy for Carbon Family Study, *Critical Reviews in Analytical Chemistry*, 46, 502–520, https://doi.org/10.1080/10408347.2016.1157013, 2016.

Appendix A: Document contents

The table of contents is listed below.

Appendix B: Acroynoms

Table B1 includes pervasive acronyms used in multiple sections.

Table B1. List of acronyms and their definitions.

Type	Acronym	Definition
Measurements	FT-IR	Fourier transform infrared
	OM	organic matter
	PM	particulate matter
	TOR	thermal optical reflectance
	OC	organic carbon
	EC	elemental carbon
	MDL	minimum detection limit
	PTFE	Polytetrafluoroethylene (Teflon)
	IMPROVE	Interagency Monitoring of PROtected Visual Environments
	CSN	Chemical Speciation Network
Site abbrev.	BYIS	Baengnyeong Island, S. Korea (IMPROVE)
	ELLA	Elizabeth, NJ (CSN)
	FRES	Fresno, CA (IMPROVE)
Chemometrics	PLS	partial least squares
	LV	latent variable
	RMSE	root mean square error
	BMCUVE	backward Monte Carlo unimportant variable elimination

Appendix C: Elements of model building and evaluation

A brief summary of model elements are shown in Table C1.

Table C1. Model elements and their descriptions.

Type	Element	Description
Data	calibration	used for model estimation
	test	used for model evaluation and performance benchmarking
	prediction	new samples to which model is to be applied
Model (PLS)	physical variables	wavenumbers
	latent variables	PLS components
	estimation	NIPALS, SIMPLS, kernel PLS, or other training algorithm
	parameter selection	CV or bootstrap using calibration samples
	spectra preparation	baseline correction or wavenumber reduction
	overall evaluation	figures of merit
	systematic evaluation	diagnostic plots: dependence of errors on concentration, site/season
	interpretation	understand most important physical and latent variables; influential samples