Manuscript prepared for Atmos. Meas. Tech. Discuss.
with version 2015/04/24 7.83 Copernicus papers of the LaTeX class copernicus.cls.
Date: 11 January 2016

Predicting ambient aerosol thermal-optical reflectance (TOR) measurements from infrared spectra: extending the predictions to different years and different sites

M. Reggente[1], A. M. Dillner[2], and S. Takahama[1]

[1]Ecole Polytechnique Federale de Lausanne, Lausanne, Switzerland
[2]University of California, Davis, California, USA

Correspondence to: S. Takahama (satoshi.takahama@epfl.ch)
and A. M. Dillner (amdillner@ucdavis.edu)

Abstract

Organic carbon (OC) and elemental carbon (EC) are major components of atmospheric particulate matter (PM), which has been associated with increased morbidity and mortality, climate change and reduced visibility. Typically OC and EC concentrations are measured using thermal optical methods such as thermal optical reflectance (TOR) from samples collected on quartz filters. In this work, we estimate TOR OC and EC using Fourier transform infrared (FT-IR) absorbance spectra from polytetrafluoroethylene (PTFE Teflon) filters using partial least square regression (PLSR) calibrated to TOR OC and EC measurements for a wide range of samples. The proposed method can be integrated with analysis of routinely collected PTFE filter samples that, in addition to OC and EC concentrations, can concurrently provide information regarding the functional group composition of the organic aerosol. We have used the FT-IR absorbance spectra and TOR OC and EC concentrations collected in the Interagency Monitoring of PROtected Visual Environment (IMPROVE) network (USA). We used 526 samples collected in 2011 at seven sites to calibrate the models, and more than 2000 samples collected in 2013 at 17 sites to test the models. Samples from six sites are present both in the calibration and test sets. The calibrations produce accurate predictions both for samples collected at the same six sites present in the calibration set ($R^2 = 0.97$ and $R^2 = 0.95$ for OC and EC respectively), and for samples from nine of the 11 sites not included in the calibration set ($R^2 = 0.96$ and $R^2 = 0.91$ for OC and EC respectively). Samples collected at the other two sites require a different calibration model to achieve accurate predictions. We also propose a method to anticipate the prediction error: we calculate the squared Mahalanobis distance in the feature space (scores determined by PLSR) between new spectra and spectra in the calibration set. The squared Mahalanobis distance provides a crude method for assessing the magnitude of mean error when applying a calibration model to a new set of samples.

1 Introduction

Organic carbon (OC) and elemental carbon (EC) are major components of atmospheric particulate matter (PM), which has been associated with increased morbidity and mortality (Janssen et al., 2011; Anderson et al., 2012), climate change (Yu et al., 2006; Bond et al., 2013) and reduced visibility (Watson, 2002; Hand et al., 2012). The major sources of EC are incomplete burning of fossil fuels and pyrolysis of biological material during combustion (Bond et al., 2007; Szidat et al., 2009). OC may be either directly emitted from sources, (e.g. incomplete combustion of organic materials) or produced from chemical reactions involving organic carbon gases (Jacobson et al., 2000). Therefore, OC and EC are monitored over long periods of time by large monitoring networks such as: the Interagency Monitoring of PROtected Visual Environments network (IMPROVE, Hand et al., 2012; Malm et al., 1994) in rural areas in the US; the Chemical Speciation Network/Speciation Trends Network (CSN/STN, Flanagan et al., 2006) in urban and suburban areas in the US; the European Monitoring and Evaluation Programme (EMEP, Tørseth et al., 2012) in Europe.

Typically OC and EC concentrations are measured using thermal methods such as thermal-optical reflectance (TOR, Chow et al., 2007; NIOSH 5040, Birch and Cary, 1996; EUSAAR-2, Cavalli et al., 2010) from samples collected on quartz filters. OC and EC are operationally defined by the temperature gradient and gaseous environment in which carbon evolves from the sample. An optical method such as reflectance or transmittance is used to correct for pyrolysis of the organic material (Cavalli et al., 2010; Chow et al., 2007). However, TOR measurements are destructive and relatively expensive.

To reduce the operating costs of large air quality monitoring networks, Dillner and Takahama (2015a, b) proposed using Fourier transform infrared spectroscopy (FT-IR) as an alternative for quantification of OC and EC. This analysis technique is inexpensive, non-destructive, and rapid. It uses PTFE samples, which are commonly used in PM monitoring networks for gravimetric mass and elemental analysis. Moreover, many quantities of interest (e.g. organic functional groups, OM and OM/OC) can be quantified from the same FT-IR spectra (Ruthenburg et al., 2014; Takahama et al., 2013; Russell, 2003).

In this work, we further evaluate the use of FT-IR as an alternate method for quantification of TOR OC and EC by extending the work of Dillner and Takahama (2015a, b) to a different year and different sites. We used FT-IR absorbance spectra and TOR OC and TOR EC measurements collected in the IMPROVE network. In the previous works, the authors used PTFE samples collected at seven sites in 2011. The filters were collected every third day; two-thirds of samples were used to calibrate the models and the remaining to test the calibration models. We used the same calibration dataset (collected in 2011) used by (Dillner and Takahama, 2015a, b), and we extended the previous analysis by: (i) evaluating the models using test samples collected at the same sites of the calibration dataset but a different year (2013); (ii) evaluating the models using test samples collected at 11 different sites and different year (2013) of the calibration dataset; (iii) proposing a statistical method to anticipate the prediction errors for each site.

To predict OC and EC values, we used partial least square regression (PLSR, Sect. 2.3). PLSR is a common method to develop calibration models for different compounds (Madari et al., 2005; Weakley et al., 2014; Vongsvivut et al., 2012). Moreover, we also propose a statistical modelling technique to anticipate the estimation error and the goodness of the estimation (Sect. 2.4). We use the squared Mahalanobis distance (Mahalanobis, 1936; Cios et al., 1998) in the feature space (scores of the PLSR; Barker and Rayens, 2003) to discriminate between sites that likely have predictions outside the predefined error range (presumably because their features are not well modelled with the available dataset).

2 Methods

2.1 IMPROVE network samples

In this work, we used samples from the IMPROVE network collected in 2011 and 2013 (Table 1 and Fig. 1). We divided these samples in four datasets (Table 1). Calibration 2011 is composed of 526 PTFE ambient samples collected at seven sites in the US in 2011 (filled and empty squares symbols in Fig. 1) plus 36 laboratory blank samples. This dataset is the same calibration set used by Dillner and Takahama (2015a, b) including Sect. S3 in the Supplement of Dillner and Takahama (2015b). Test 2011 is composed of 269 PTFE ambient samples (plus 18 PTFE laboratory blank samples) collected at the same sites and the same year (2011) of the calibration dataset (filled and empty squares symbols in Fig. 1). This dataset is the same test set used by Dillner and Takahama (2015a, b). Test 2013 is composed of 949 PTFE ambient samples (plus 50 PTFE laboratory blank samples) collected at six sites that are the same sites but different year (2013) of the calibration dataset (filled squares symbols in Fig. 1). Test 2013 additional (Addl) is composed of 1290 PTFE ambient samples (plus the same 50 PTFE laboratory blank samples of the Test 2013 dataset) collected at 11 different sites and different year (2013) of the calibration dataset (black triangles in Fig. 1). Four of the eleven additional sites are urban sites, some of the samples experienced significant smoke impact and some were collected at an IMPROVE site in South Korea (with high sample loadings).

The IMPROVE network collects samples every third day from midnight to midnight (local time) at a nominal flow rate of 22.8 $\mathrm{L\,min^{-1}}$, which yields a nominal volume of 32.8 $\mathrm{m^3}$ and produces samples of particles smaller than 2.5 µm in diameter ($PM_{2.5}$). The FT-IR analysis is applied to 25 mm PTFE samples (Teflon, Pall Gelman – 3.53 $\mathrm{cm^2}$ sample area) that are analyzed for gravimetric mass, elements and light absorption in the IMPROVE network. Quartz filters collected in parallel to the PTFE samples are analyzed by TOR using the IMPROVE A protocol to obtain OC and EC mass in the IMPROVE network (Chow et al., 2007). The OC and EC values are also adjusted to account for flow differences between the quartz and PTFE samples. IMPROVE samples lacking either flow records for PTFE filters or TOR measurements are excluded.

2.2 FT-IR analysis: spectra acquisition

We analyzed a total of 3034 PTFE ambient samples and 104 PTFE laboratory blank samples using a Tensor 27 Fourier transform infrared (FT-IR) spectrometer (Bruker Optics, Billerica, MA) equipped with a liquid-nitrogen-cooled wide-band mercury cadmium telluride detector. The samples are analyzed using transmission FT-IR in air that has low levels of water vapor and CO_2 using an empty sample compartment as the reference (a more detailed description is in Dillner and Takahama, 2015a). The filter samples are not treated prior to FT-IR analysis except that values interpolated during the zero-filling process are removed. These spectra contain 2784 wavenumbers.

2.3 Building the calibration models

We used the calibration models developed by Dillner and Takahama (2015a, b) to predict TOR OC and EC in the 2013 data sets. For each site, the samples collected in 2011 are ordered by date and every third sample is removed and included in the Test 2011 dataset. The remaining samples are placed in the Calibration 2011 dataset.

Briefly, the calibraton models were developed using Partial Least Squares Regression (PLSR, Wold et al., 1983; Geladi and Kowalski, 1986) using the kernel PLS algorithm, implemented by the PLS library (Mevik and Wehrens, 2007) of the R statistical package (R Core Team, 2014). The goal is to predict a set of coefficients $\boldsymbol{b}$ from a matrix of spectra $\mathbf{X}$ for observation $\boldsymbol{y}$ (OC or EC), with residuals $\boldsymbol{e}$:

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{b} + \boldsymbol{e} \tag{1}$$

PLSR circumvents issues that arise when collinearity exists among variables in $\mathbf{X}$ (strong correlation of absorbances across wavenumbers) and when the number of variables (wavenumbers) in the spectra matrix $\mathbf{X}$ exceeds the number of observation (rows of $\mathbf{X}$). PLSR performs a bilinear decomposition of both $\mathbf{X}$ and $\boldsymbol{y}$: the matrix of spectra ($\mathbf{X}$) is decomposed into a product of orthogonal factors (loadings, $\mathbf{P}$) and their respective contributions (scores, $\mathbf{T}$). Observed variations in the OC or EC mass are reconstructed through a combination of these factors ($\mathbf{T}$) and a set of weights simultaneously ($\boldsymbol{q}$) developed to relate features to the dependent and independent variables – scores and loadings describe the covariance between $\mathbf{X}$ and $\boldsymbol{y}$.

$$\mathbf{X} = \mathbf{T}\mathbf{P}^\top + \mathbf{E} \tag{2}$$

$$\boldsymbol{y} = \mathbf{T}\boldsymbol{q}^\top + \boldsymbol{f} \tag{3}$$

The two sets of factors are related through a weighting matrix **W** to reconstruct the set of regression coefficient $\boldsymbol{b}$:

$$\boldsymbol{b} = \mathbf{W}\left(\mathbf{P}^{\top}\mathbf{W}\right)^{-1}\boldsymbol{q}^{\top} \tag{4}$$

Candidate models for calibration are generated by varying the number of factors used to represent the matrix of spectra. To select the number of factors we used $K = 10$ fold cross validation (CV, Hastie et al., 2009; Arlot and Celisse, 2010) on the calibration 2011 dataset (Sect. 2.1). We used the minimum root mean square error (RMSEP) to select the model with least prediction error.

For the calibration of the FT-IR spectra to TOR EC, in order to eliminate bias in the calibration and improve prediction capability for low EC samples (EC $< 2.4\,\mu$g), we used the hybrid calibration approach described in (Dillner and Takahama, 2015b). We used two calibration models: the first uses samples in the calibration set that are in the lowest one-third of the EC mass range to predict samples in the test set that are also in the lowest one-third of the EC mass range; the second one for the remaining samples. Localization of the calibration (with respect to concentration) is a commonly used method to improve the performance of the calibration, often at the more difficult to measure low end of the range.

2.4 Anticipating prediction errors

For samples collected during different periods or different locations, it is useful to know whether the present calibration model is appropriate for a new sample or a new set of samples. Using features present in FT-IR spectra, we propose a method to anticipate the prediction error in OC or EC concentrations prior to applying the calibration model. The purpose of such an approach is to determine whether a particular calibration model is suitable for a new set of samples without requiring an assessment of prediction accuracy using TOR OC and EC measurements a posteriori. The feature space is a low-dimensional projection of the absorbance spectra that has been associated with prediction capability for TOR OC and EC, and is determined by the factor scores determined by PLSR (Eqs. 2 and 3). We calculate the centroid ($\boldsymbol{\mu}$) and the covariance (Σ) of the calibration samples projected in the feature space:

$$\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_k)^\top, \quad \mu_j = \frac{1}{n}\sum_{i=1}^{n} t_{ij} \quad \forall j = 1, \ldots, k \tag{5}$$

$$\Sigma = \mathrm{cov}\left(\boldsymbol{t}_i, \boldsymbol{t}_j\right), \quad i = 1 \ldots, n \quad j = 1 \ldots, k \tag{6}$$

where k is the number of factors used to represent the matrix of spectra; n is the number of PTFE samples included in the Calibration 2011 dataset; $\boldsymbol{t}$ are the columns of the scores matrix.

For each calibration sample, we calculate the squared Mahalanobis distance (D^2_{M}) between the sample itself and the centroid of the calibration set, taking into account the covariance matrix to normalize the distance according to the magnitude of dispersion in each dimension.

$$D^2_{\mathrm{M},i} = (\boldsymbol{t}_i - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{t}_i - \boldsymbol{\mu}), \quad i = 1, \ldots, n \tag{7}$$

We then project absorbance spectra of the test set (Table 1) in the feature space, and calculate the D^2_{M} between each test sample and the centroid ($\boldsymbol{\mu}$) of the calibration set.

We present our evaluation for two cases. In the first case, we calculate the D^2_{M} and errors for the calibration and test sets according to sampling site, and calculated the mean D^2_{M} and mean absolute error for each site. In the second case, we considered the D^2_{M} and absolute error for each sample, without aggregation. We considered acceptable predictions, the ones that have errors within the magnitude of the 2011 dataset, and unacceptable predictions those for which the errors are greater. As boundaries for the discrimination, in the first case, we used the greatest mean D^2_{M} plus one standard error (SE) and the greatest mean absolute error plus one SE found in the Test 2011 set. In the second case, we used the greatest D^2_{M} and the absolute error found in the Test 2011 set (except for one sample that we considered an outlier, Fig. S3 in the Supplement).

The samples collected in 2013 are divided into four classifications, which can be visualized by membership in one of four quadrants in a plot of absolute error vs. D^2_{M} (Figs. 5 and

11). True negative (TN) samples (or sites) are those for which the D^2_{M} (or mean D^2_{M}) and absolute error (or mean absolute error) fall in the third (bottom left) quadrant. For TN samples, the D^2_{M} gives a reliable indication that the prediction error is within magnitude of 2011 samples. True positive (TP) samples (or sites) are those that lie in the first (upper right) quadrant; the D^2_{M} provides a reliable indication that the prediction error is greater than the magnitude of 2011 samples. False negative (FN) samples (or sites) lie in the second (upper left) quadrant; the D^2_{M} is not indicative of the increased errors above the 2011 predictions. False positive (FP) samples (or sites) are those that lie in the fourth (bottom right) quadrant; sample spectra have significantly higher D^2_{M}s but do not have increased prediction errors over 2011 predictions.

2.5 Model evaluation

We evaluated the calibration models (trained with the calibration 2011 dataset, Sect. 2.3) on the three datasets described in Sect. 2.1. The quality of each calibration is evaluated by calculating four performance metrics: bias, error, normalized error and the coefficient of determination (R^2) of the linear regression fit of the predicted FT-IR OC and FT-IR EC to measured TOR OC and EC. FT-IR OC and EC are the OC and EC predicted from the FT-IR spectra and the PLSR calibration model. The bias is the median difference between measured (TOR) and predicted (FT-IR) for the test set. Error is the median absolute bias. The normalized error for a single prediction is the error divided by the TOR value. The median normalized error is reported to provide an aggregated estimate. The performance metrics are also calculated for the collocated TOR observations and compared to those of the FT-IR OC and FT-IR EC to TOR OC and TOR EC regression. The minimum detection limit (MDL) and precision of the FT-IR and TOR methods are calculated and compared. The MDL of the FT-IR method is 3 times the standard deviation of the laboratory blank samples in the test sets (Sect. 2.1). The MDL for the TOR method is 3 times the standard deviation of 514 blank samples (Desert Research Institute, 2012). Precision for both FTIR and TOR is calculated using 14 parallel samples in the dataset Test 2011 at the Phoenix site; 240 parallel samples in the dataset Test 2013 at the Phoenix and Proctor Maple R. F. sites; and

115 parallel samples in the dataset Test 2013 Addl at the Yosemite site. For evaluation of the FT-IR predictions against TOR reference values, we used 621 measurements collected in 2013 from seven IMPROVE sites with collocated TOR measurements (Everglades, Florida; Hercules Glade, Missouri; Hoover, California; Medicine Lake, Montana; Phoenix, Arizona; Saguaro West, Arizona; Seney, Virginia).

3 Results and Discussion

In this section, we first evaluate and discuss the performance of the models by comparing the predicted with the observed TOR OC (Sect. 3.1) and TOR EC (Sect. 3.3) measurements for ambient samples collected at the four datasets described in Sect. 2.1. Then, in Sects. 3.2 and 3.4 we describe the results of the prediction error anticipation for the FT-IR OC and FT-IR EC respectively.

3.1 Prediction of TOR OC from FT-IR spectra

The comparison between predicted FT-IR OC and measured TOR OC for the datasets described in Sect. 2.1 is shown in Figure 2. The first row refers to the Calibration 2011 and Test 2011 dataset. In this work, the results obtained with the Test 2011 are used for comparison with the results obtained with the Test 2013 and Test 2013 Addl datasets. The detailed description and discussion of the Test 2011 results can be found in (Dillner and Takahama, 2015a).

In the case of predictions based on ambient sample collected at the same sites of the calibration dataset but in a different year (Test 2013 dataset, bottom left panel in Fig. 2), we observe that the performance metrics show good agreement between measured (TOR OC) and predicted OC values (FT-IR OC). Moreover, comparing these metrics with the ones obtained with the Test 2011 dataset, we note that the R^2 is better in the former, and the bias, error and normalized error are slightly worse in the former. However, these differences are not substantial, and we can conclude that the predictions made with the Test 2011 and Test 2013 datasets are similar. A T test at the confidence level of 95 % between the

Test 2011 and Test 2013 predictions, and one between their absolute errors confirmed this observation. From Fig. 3 (performance of the collocated TOR OC measurements), we observe that the metrics are similar to the ones obtained with the Test 2011 dataset. The bias (0.09 µg m^{-3}), error (0.12 µg m^{-3}) and normalized error (18 %) of the Test 2013 dataset are slightly worse than the ones obtained with the collocated TOR OC (0 µg m^{-3}, 0.06 µg m^{-3} and 11 % respectively). However, these differences are not substantial as shown in Table 2, which compares the MDLs and precisions of FT-IR OC predictions and TOR OC measurements. Table 2 shows that both the MDL and precision of the Test 2013 are in the same range of the TOR OC and slightly better than the ones obtained with the Test 2011 dataset. It is also interesting to note that in the Test 2013 dataset there are two samples notably above the maximum values used in the calibration dataset, and the model is still able to predict them accurately. This observation agrees with the results found by Dillner and Takahama (2015a) (Non-uniform A case), in which the authors used samples with TOR OC in the lowest two-thirds of the TOR OC range to predict samples with TOR OC in the highest one-third of the TOR OC range, and the highest one-third of samples were well predicted.

The bottom right panel in Fig. 2 shows the performance of the calibration model for samples collected at different sites and different year (Test 2013 Addl dataset) than the samples used for calibration. The bias, error and normalized error (0.06 µg m^{-3}, 0.13 µg m^{-3} and 14 %) are in between the one found for the Test 2011 and Test 2013 datasets. However, the scatterplot and the R^2 metric show worse agreement between measured TOR OC and predicted FT-IR OC values ($R^2 = 0.89$), than the ones obtained with the Test 2011 and Test 2013 datasets ($R^2 = 0.94$ and 0.97 respectively). The scatter plot also shows that the model tends to overestimate the OC values. A T test at the confidence level of 95 % between the Test 2013 and Test 2013 Addl predictions, and one between their absolute errors shows a statistical difference between the results obtained with the two datasets. These results can be explained by the consideration that the sites where the samples of the Test 2013 Addl were collected may have different composition and loadings of OC than the ones used in the calibration dataset. Therefore, the calibration model does not have enough information to model the different features of these sites. However, the MDL and precision are similar to the one obtained with the collocated TOR dataset, indicating that low concentrations are more likely to be precise than high concentrations. This could be explained by the fact that the majority of the ambient samples used for calibration were collected at rural sites (76 % from six sites), and the ones collected at the urban site (Phoenix) may have different features (composition and mass range) from the ones collected in the Test 2013 Addl dataset, many of which are urban or highly polluted (Birmingham (Alabama), Fresno (California), Puget Sound (Seattle, Washington) and South Korea).

This observation is supported by the results in Fig. 4. In this case, the calibration and the test datasets are composed of the ambient samples collected at the rural sites of each dataset. The performance metrics show accurate agreement between measured (TOR OC) and predicted OC values (FT-IR OC) for all the datasets. Moreover, it is worth noting that in the Test 2013 and Test 2013 Addl there are samples that are above 200 µg (probably due to fire events). Even though these samples are outside the calibration range, the model is still able to predict them with reasonable accuracy (with a tendency of overestimation).

On the basis of these results, we can conclude that: (i) it is possible to use PTFE ambient samples, to calibrate a model that predicts accurately TOR OC values from PTFE ambient samples collected at the same sites (both rural and urban) in the same or different years of the ones used for the calibration; (ii) it is possible to use PTFE ambient samples, collected at rural sites, to calibrate a model that predicts accurately TOR OC values from PTFE ambient samples collected at different rural sites and in different years of the ones used for the calibration.

3.2 Anticipation of the prediction error: FT-IR OC

As shown in Fig. 5, the D^2_{M}s are smaller in the calibration set than in the test set because the latent variables comprising the feature space were defined using the former. Moreover, the D^2_{M}s increase for the test samples that have different features from the ones used in calibration. Figure 5 shows the aggregated (per site) mean D^2_{M} against the mean absolute error for each dataset. In accordance with the results shown in Sect. 3.1, the Test 2011 and Test 2013 sites present D^2_{M}s similar to the ones found in the calibration dataset (all sites lie in TN sector). On the other hand, the Test 2013 Addl dataset contains two sites (10 represents the South Korea site and 11 represents the Fresno site, Table 1) which have both mean D^2_{M} and mean absolute errors above the boundaries used for discrimination of unacceptable prediction errors (TP sector). This diagnostic indicates that the two sites may contain different sources and chemical composition that are not well represented in the Calibration 2011 dataset.

The scatter plot and the performance metrics of the Test 2013 Addl without the samples collected at the two sites anticipated (and confirmed) to have high errors are shown in Fig. 6. The R^2 metric notably improves from 0.89 to 0.96, and the remaining evaluation statistics are also improved. Moreover, a T test at the confidence level of 95 % between the predictions and one between their absolute errors shows a statistical difference between the results obtained with the two datasets.

The evaluation of predictions using a calibration model constructed from only the Korea and Fresno sites is shown in Fig. 7. The calibration set uses two-thirds of the ambient samples (we followed the same methodology that we have used to prepare the Calibration 2011 dataset) collected in 2013 at the two sites and one third for test. The results show that with the appropriate calibration samples, we can achieve accurate predictions of TOR OC values (bias $= -0.03\,\mu\mathrm{g\,m^{-3}}$, error $= 0.16\,\mu\mathrm{g\,m^{-3}}$, normalized error $= 10\,\%$ and $R^2 = 0.96$) also in these two sites. For comparison, Fig. S1 shows the evaluation of predictions at the Korean and Fresno sites using the Calibration 2011 dataset (bias $= 0.28\,\mu\mathrm{g\,m^{-3}}$, error $= 0.43\,\mu\mathrm{g\,m^{-3}}$, normalized error $= 25\,\%$ and $R^2 = 0.79$). A T test at the confidence level of 95 % between the predictions and one between their absolute errors shows a statistically significant reduction in mean errors for predictions using the new calibration (Fig. 7) and those using the base case calibration (Fig. S1). Therefore, we can conclude that sites with samples that are on average dissimilar to those in the calibration are shown to benefit from the construction of a separate calibration model.

The results of the D^2_{M} against the absolute error for each sample (without any aggregation) are reported in Tab. 3. In the Supplement, we show the plots of the D^2_{M} distances against the absolute errors for each site (Figs. S2–S5). For each dataset, Table 3 reports both the percentage of samples falling in the FN, FP, TN and TP sectors and the performance metrics of the models that exclude unacceptable predictions (samples falling in the FP and TP sectors). The Test 2013 dataset presents a low percentage (0.4 %) of erroneous classifications (samples falling in the FN and FP sectors), and the performance metrics are in line with what we found in Fig. 2.

The Test 2013 Addl contains 3.2 % of well classified unacceptable predictions TP. Most of these unacceptable predictions are due to samples collected at the Korean (1.9 %) and Fresno (0.9 %) sites (0.4 % from other 3 sites, Fig. S5). Moreover, the Test 2013 Addl contains 2.3 % of erroneous classifications falling in either FP or FN sectors (1.1 % is from the Korean samples, 0.4 % from Fresno samples and 0.8 % from other 7 sites, Fig. S5). The R^2 metric improves from 0.89 to 0.92 with respect to the results found in Fig. 2. The bias, error and normalized error are similar. However, because 1.7 % of samples are in the FN sector (and we consider them acceptable), this explains the lower prediction performance in comparison to the case in which all the Korean and Fresno samples are excluded (Fig. 6).

This analysis suggests that spectral signals projected into the feature space of a particular PLS calibration model contain useful information for anticipating the magnitude of prediction errors. The ability to anticipate the quality of predictions based on spectra features is relevant for strategic collection of calibration samples and selection of available samples from which a calibration model can be constructed. We expect that the capability for discrimination at the level of individual samples can be improved in future studies. The D^2_{M} is a strong classifier in the case that the scores $\mathbf{T}$ are normally distributed in the multivariate (MV, in this case 47 dimensions) feature space. Indeed, the D^2_{M} (Eq. 7) is the exponential term of the MV Normal (MVN) distribution and represents the distance between each point and the mean of the MVN distribution. Assessing the assumption of multivariate normality is a challenging process because different methods may provide different results under different assumptions and conditions (Mecklin and Mundfrom, 2005). Both the Henze–Zirkler (Henze and Zirkler, 1990) and Mardia MV Normality test (Mardia, 1970) on the calibration scores lead to a rejection of the null hypothesis (at 95 % confidence level) that the scores are Normally distributed. It is plausible to conceive of spectral preprocessing methods to improve the MVN assumption, or invoke different similarity metrics that do not require a specific distribution of points in the feature space. However, even while the assumption of MVN is not fulfilled, we report that the mean D^2_{M} provides indication of samples dissimilar to those comprising the calibration set when aggregated at the site level.

3.3 Prediction of TOR EC from FT-IR spectra

In this section, we extend the analysis done in Sects. 3.1 and 3.2 to the case of TOR EC measurements. For this case, as described in Sect. 2.3, we use the hybrid model proposed by Dillner and Takahama (2015b).

The comparison between predicted FT-IR EC and measured TOR EC for the datasets described in Sect. 2.1 is shown in Fig. 8. The first row refers to the Calibration 2011 and Test 2011 dataset. In this work, the results obtained with the Test 2011 are used for comparison with the results obtained with the Test 2013 and Test 2013 Addl datasets. The detailed description and discussion of the results obtained with the Test 2011 dataset can be found in (Dillner and Takahama, 2015b).

In the case of predictions based on ambient sample collected at the same sites of the calibration dataset but different year (Test 2013 dataset, bottom left panel in Fig. 8), we observe that the performance metrics show good agreement between measured (TOR EC) and predicted EC values (FT-IR EC). Moreover, it shows similar performance to the prediction based on ambient samples collected at the same sites and year of the calibration (Test 2011). A T test at the confidence level of 95 % between the Test 2011 and Test 2013 predictions, and one between their absolute errors confirmed this observation. The results are similar to FT-IR OC (Sect. 3.1).

From Fig. 9 (performance of the collocated TOR EC measurements), we observe that the metrics are similar to the ones obtained with the Test 2011 and Test 2013 datasets. The normalized error is higher for FT-IR EC (21 and 24 % for Test 2011 and Test 2013 respectively) than for collocated TOR EC measurements (14 %). Table 4 compares the MDLs and precisions of FT-IR EC predictions and TOR EC measurements and it shows that the MDL is very similar to TOR EC and the precision is better than (less than half) of TOR EC.

The bottom right panel in Fig. 8 shows the performance of the calibration model for samples collected at different sites and different year (Test 2013 Addl dataset) of the samples used for calibration. We observe that the performance metrics show worse agreement between measured TOR EC and predicted EC values (FT-IR EC), than the ones obtained with the Test 2011 and Test 2013 datasets. A T test at the confidence level of 95 % between the Test 2013 and Test 2013 Addl predictions, and one between the absolute errors show a statistical difference between the results obtained with the two datasets.

For the case of OC, we found that the model, based on rural sites only, predicts accurately TOR OC values from PTFE ambient samples collected at different rural sites and different years of the ones used for the calibration (Sect. 3.1). For the case of EC, the models trained only with the rural sites (Fig. 10) have lower R^2 (0.88, 0.90 and 0.87 respectively for the three test sets) than the ones that use the entire calibration set (0.96, 0.95 and 0.87, Fig. 8). However, looking at the plots in Fig. 10 we observe that the predicted FT-IR EC values do not differ substantially from the TOR EC values, and the worse performance metric is explained by the low concentrations measured at the rural sites: R^2 decreases (the total sum of squares tends to zero) when the measurements are close to zero.

3.4 Anticipation of the prediction error: FT-IR EC

The aggregated (per site) mean D_{M}^2 against the mean absolute error for each dataset is shown in Fig. 11. Similar to the results shown in Sect. 3.3, the Test 2011 (except for site 5) and Test 2013 sites present D_{M}^2 similar to the ones found in the calibration dataset (all the sites lie in TN sector). Unlike the OC case (Sect. 3.2), the Birmingham site (number 8) is misclassified (it lies in the FN sector). Moreover, the site 11 (Fresno) has a mean D_{M}^2 close to the boundary, but it lies in the TP sector. The Korean site (10) has both mean D_{M}^2 and mean absolute errors above the boundaries used for discrimination between acceptable and unacceptable predictions (TP sector). Figure 12 shows the scatter plot and the performance metrics of the Test 2013 Addl without the samples collected at the two sites we have classified not acceptable (Fresno and Korea). We note that the predictions are less spread and the R^2 metric improve from 0.87 to 0.91. The remaining metrics are almost identical

(Fig. 8 for comparison). Moreover, a T test at the confidence level of 95 % between the predictions, and one between their absolute errors shows a statistical difference between the results obtained with the two datasets.

The evaluation of predictions using a calibration model constructed from only the Korea and Fresno sites is shown in Fig. 13. The calibration set uses two-thirds of the ambient samples (we followed the same methodology that we have used to prepare the Calibration 2011 dataset proposed by Dillner and Takahama, 2015a, b) collected in 2013 at the two sites and one third for test. The results show that with the appropriate dataset, also in Fresno we can achieve accurate predictions of EC values ($R^2 = 0.93$, bias $= 0$ µg m^{-3}, error $= 0.06$ µg m^{-3} and normalized error $= 11$ %). The performance metrics at the Korean site ($R^2 = 0.66$, bias $= -0.07$ µg m^{-3}, error $= 0.11$ µg m^{-3} and normalized error $= 18$ %) are not as good as Fresno and are mostly due to one sample. The predictions for all other samples are more accurate. By removing the one erroneous sample, R^2 increases from 0.66 to 0.84. For comparison, Fig. S7 shows the evaluation of predictions at the Korean and Fresno sites using the Calibration 2011 dataset ($R^2 = 0.85$, bias $= 0.05$ µg m^{-3}, error $= 0.10$ µg m^{-3} and normalized error $= 22$ % and $R^2 = 0.60$, bias $= 0.13$ µg m^{-3}, error $= 0.17$ µg m^{-3} and normalized error $= 33$ % at Fresno and Korea site respectively). A T test at the confidence level of 95 % between the predictions and one between their absolute errors shows a statistically significant reduction in mean errors for predictions using the new calibration (Fig. 13) and those using the base case calibration (Fig. S7). Therefore, we can conclude that sites with samples that are on average dissimilar to those in the calibration are shown to benefit from the construction of a separate calibration model.

The results of the D^2_{M} (and absolute error) for each sample (without any aggregation) are reported in Tab. 5, and in the Supplement we show the corresponding figures (Figs. S8–S11). For each dataset, Table 5 reports both the percentage of samples falling in the FN, FP, TN and TP sectors and the performance metrics of the models that exclude unacceptable samples (those falling in the FP and TP sectors). The Test 2013 dataset presents a low percentage (0.5 %) of erroneous classifications (samples falling in the FN and FP sectors), and the performance metrics, for the test sets with unacceptable samples excluded, are in line with what we found when all samples are included (Fig. 8). The Test 2013 Addl contains 1 % of well classified unacceptable predictions TP. Mostly of these unacceptable predictions are due to samples collected at the Korean site (0.7 %, Fresno, 0.2 %, Hoover, 0.1 %, Fig. S11). Moreover, the Test 2013 Addl presents 2.8 % of erroneous classifications (Korea, 1.2 %, Fresno, 0.7 %, four sites, 0.9 %, Fig. S11). The R^2 metric, for the test sets with unacceptable samples excluded, improves from 0.87 to 0.92 with respect to the results found when all samples are included (Fig. 8). The bias, error and normalized error are similar.

We note that for this work, the boundaries are chosen heuristically (maximum D^2_{M} and absolute error found in the 2011 dataset (Sect. 2.4)), and they tend to classify the great majority of the samples (96.3 %) in the Test 2013 Addl dataset as well predicted (TN). The choice of different boundaries, spectral preprocessing, or distance metric (as discussed for OC in Sect. 3.2), may lead to a less generous classification and an improved discrimination between acceptable and unacceptable predictions, with particular care to minimize FN classifications. However, as for OC, the analysis of EC suggests that spectral signals projected into the feature space of a particular PLS calibration model contain useful information for anticipating the magnitude of prediction errors, and using the D^2_{M} we still obtain useful results as tested by the performance metrics in Fig. 12 and Table 5. The anticipation error for the FT-IR EC predictions is less accurate than the one found in the OC case (the Birmingham site was misclassified for EC but not for OC), but we expect that the capability for discrimination can be improved in future work for the same reasons given in Sect. 3.2.

4 Conclusions

In this work, we have used FT-IR spectra of particles collected on ambient PTFE filters (without spectral preprocessing) to predict the concentrations of TOR OC and EC using partial least square regression. We extended the work of Dillner and Takahama (2015a, b) by (i) evaluating the models using test samples collected at the same sites as the calibration dataset but a different year (2013); (ii) evaluating the models using test samples collected at 11 different sites and a different year (2013) than the calibration dataset; (iii) proposing a statistical method to anticipate the prediction errors for each site.

Our results show that ambient samples collected at the same sites in the same or different years as the ones used for the calibration accurately predicts TOR OC and EC values ($0.94 \leq R^2 \leq 0.97$ for FT-IR OC and $0.95 \leq R^2 \leq 0.96$ for FT-IR EC). Moreover, the precisions and the MDLs of the FT-IR OC and FT-IR EC are in the same order of the precision and MDL of the measurements reported by the TOR methods. We further determine that using samples collected at rural sites to calibrate a model for OC improves the prediction at different rural locations and in different years as the ones used for the calibration ($R^2 = 0.97$). This result is likely due to similar loadings and OC sources at these sites. In contrast, due to the low EC concentrations (EC $\leq 60\,\mu g$) recorded at the rural sites, the EC predictions are less accurate ($R^2 = 0.87$) because of the lower signal to noise ratio.

We further demonstrate that spectral features contain information that can be used to anticipate prediction errors for OC and EC using a particular calibration model. In our assessment using the squared Mahalanobis distance in the feature space of PLSR, sites with samples which are on average dissimilar to those in the calibration set are identified and shown to benefit from the construction of separate calibration models. Building separate calibration models leads to improvements in R^2 from 0.79 to 0.96 for the OC evaluation, and from 0.60 to 0.66 and from 0.85 to 0.93 for EC at Korea and Fresno site, respectively. These improvements are statistically significant as determined by a test of significance in the mean absolute errors. We posit possible that this assessment can be improved with spectral preprocessing or additional statistical algorithms; at present time this work illustrates a framework that may guide strategic calibration model development and calibration sample collection.

Therefore, we conclude that FT-IR spectra of aerosol samples collected on PTFE filters can be calibrated to TOR measurements (using partial least-squares regression) to provide robust predictions of TOR OC and EC concentrations. The calibration models based on samples collected in selected sites can be used in subsequent years and at locations that, on average, have similar features to those in the calibration set. Moreover, identification of sites that are, on average, dissimilar to those in the calibration set provides an opportunity to develop dedicated calibration models specific to samples from these sites.

The Supplement related to this article is available online at doi:10.5194/amtd-0-1-2016-supplement.

Acknowledgements. The authors acknowledge funding from the IMPROVE program (National Park Service cooperative agreement P11AC91045) and EPFL funding. Thanks to the IMPROVE team at UC Davis for performing the sample handling and site maintenance for all IMPROVE sites.

References

Anderson, J., Thundiyil, J., and Stolbach, A.: Clearing the Air: A Review of the Effects of Particulate Matter Air Pollution on Human Health, Journal of Medical Toxicology, 8, 166–175, doi:10.1007/s13181-011-0203-1, 2012.

Arlot, S. and Celisse, A.: A survey of cross-validation procedures for model selection, Statist. Surv., 4, 40–79, doi:10.1214/09-SS054, 2010.

Barker, M. and Rayens, W.: Partial least squares for discrimination, J. Chemometr., 17, 166–173, 2003.

Birch, M. E. and Cary, R. A.: Elemental carbon-based method for occupational monitoring of particulate diesel exhaust: methodology and exposure issues, Analyst, 121, 1183–1190, 1996.

Bond, T. C., Bhardwaj, E., Dong, R., Jogani, R., Jung, S., Roden, C., Streets, D. G., and Trautmann, N. M.: Historical emissions of black and organic carbon aerosol from energy-related combustion, 1850–2000, Global Biogeochem. Cy., 21, GB2018, doi:10.1029/2006GB002840, 2007.

Bond, T. C., Doherty, S. J., Fahey, D. W., Forster, P. M., Berntsen, T., DeAngelo, B. J., Flanner, M. G., Ghan, S., Kärcher, B., Koch, D., Kinne, S., Kondo, Y., Quinn, P. K., Sarofim, M. C., Schultz, M. G., Schulz, M., Venkataraman, C., Zhang, H., Zhang, S., Bellouin, N., Guttikunda, S. K., Hopke, P. K., Jacobson, M. Z., Kaiser, J. W., Klimont, Z., Lohmann, U., Schwarz, J. P., Shindell, D., Storelvmo, T., Warren, S. G., and Zender, C. S.: Bounding the role of black carbon in the climate system: A scientific assessment, J. Geophys. Res.-Atmos., 118, 5380–5552, doi:10.1002/jgrd.50171, 2013.

Cavalli, F., Viana, M., Yttri, K. E., Genberg, J., and Putaud, J.-P.: Toward a standardised thermal-optical protocol for measuring atmospheric organic and elemental carbon: the EUSAAR protocol, Atmos. Meas. Tech., 3, 79–89, doi:10.5194/amt-3-79-2010, 2010.

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, JAPCA J. Air Waste Ma., 57, 1014–1023, doi:10.3155/1047-3289.57.9.1014, 2007.

Cios, K., Pedrycz, W., and Swiniarski, R. W.: Data Mining Methods for Knowledge Discovery, Kluwer Academic Publishers, Norwell, MA, USA, 1998.

Dillner, A. M. and Takahama, S.: Predicting ambient aerosol thermal-optical reflectance (TOR) measurements from infrared spectra: organic carbon, Atmos. Meas. Tech., 8, 1097–1109, doi: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, Atmos. Meas. Tech., 8, 4013–4023, doi:10.5194/amt-8-4013-2015, 2015b.

Flanagan, J. B., Jayanty, R., Rickman, Jr, E. E., and Peterson, M. R.: $PM_{2.5}$ speciation trends network: evaluation of whole-system uncertainties using data from sites with collocated samplers, JAPCA J. Air Waste Ma., 56, 492–499, 2006.

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

Hand, J., Schichtel, B., Pitchford, M., Malm, W., and Frank, N.: Seasonal composition of remote and urban fine particulate matter in the United States, J. Geophys. Res.-Atmos., 117, D05209, doi:10.1029/2011JD017122, 2012.

Hastie, T. J., Tibshirani, R. J., and Friedman, J. H.: The elements of statistical learning : data mining, inference, and prediction, Springer series in statistics, Springer, New York, 2009.

Henze, N. and Zirkler, B.: A class of invariant consistent tests for multivariate normality, Commun. Stat.-Theory M., 19, 3595–3617, 1990.

Jacobson, M., Hansson, H., Noone, K., and Charlson, R.: Organic atmospheric aerosols: review and state of the science, Reviews of Geophysics, 38, 267–294, 2000.

Janssen, N., Hoek, G., Simic-Lawson, M., Fischer, P., Van Bree, L., Ten Brink, H., Keuken, M., Atkinson, R. W., Anderson, H. R., Brunekreef, B., and Cassee, F. R.: Black carbon as an additional indicator of the adverse health effects of airborne particles compared with PM_{10} and $PM_{2.5}$, Environ. Health Persp., 119, 1691–1699, 2011.

Madari, B. E., Reeves III, J. B., Coelho, M. R., Machado, P. L., De-Polli, H., Coelho, R. M., Benites, V. M., Souza, L. F., and McCarty, G. W.: Mid-and near-infrared spectroscopic determination of carbon in a diverse set of soils from the Brazilian national soil collection, Spectrosc. Lett., 38, 721–740, 2005.

Mahalanobis, P. C.: On the generalized distance in statistics, Proceedings of the National Institute of Sciences (Calcutta), 2, 49–55, 1936.

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

Mardia, K. V.: Measures of multivariate skewness and kurtosis with applications, Biometrika, 57, 519–530, 1970.

Mecklin, C. J. and Mundfrom, D. J.: A Monte Carlo comparison of the Type I and Type II error rates of tests of multivariate normality, J. Stat. Comput. Sim., 75, 93–107, 2005.

Mevik, B.-H., and Wehrens, R.: The pls package: principal component and partial least squares regression in R, Journal of Statistical Software, 18, 1–24, 2007.

R Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2014.

Russell, L. M.: Aerosol organic-mass-to-organic-carbon ratio measurements, Environ. Sci. Technol., 37, 2982–2987, 2003.

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

Szidat, S., Ruff, M., Perron, N., Wacker, L., Synal, H.-A., Hallquist, M., Shannigrahi, A. S., Yttri, K. E., Dye, C., and Simpson, D.: Fossil and non-fossil sources of organic carbon (OC) and elemental carbon (EC) in Göteborg, Sweden, Atmos. Chem. Phys., 9, 1521–1535, doi:10.5194/acp-9-1521-2009, 2009.

Takahama, S., Johnson, A., and Russell, L. M.: Quantification of carboxylic and carbonyl functional groups in organic aerosol infrared absorbance spectra, Aerosol Sci. Tech., 47, 310–325, 2013.

Tørseth, K., Aas, W., Breivik, K., Fjæraa, A. M., Fiebig, M., Hjellbrekke, A. G., Lund Myhre, C., Solberg, S., and Yttri, K. E.: Introduction to the European Monitoring and Evaluation Programme (EMEP) and observed atmospheric composition change during 1972–2009, Atmos. Chem. Phys., 12, 5447–5481, doi:10.5194/acp-12-5447-2012, 2012.

Vongsvivut, J., Heraud, P., Zhang, W., Kralovec, J. A., McNaughton, D., and Barrow, C. J.: Quantitative determination of fatty acid compositions in micro-encapsulated fish-oil supplements using Fourier transform infrared (FTIR) spectroscopy, Food Chem., 135, 603–609, 2012.

Watson, J. G.: Visibility: science and regulation, JAPCA J. Air Waste Ma., 52, 628–713, 2002.

Weakley, A. T., Miller, A. L., Griffiths, P. R., and Bayman, S. J.: Quantifying silica in filter-deposited mine dusts using infrared spectra and partial least squares regression, Anal. Bioanal. Chem., 406, 4715–4724, 2014.

Wold, S., Martens, H., and Wold, H.: The multivariate calibration problem in chemistry solved by the PLS method, in: Matrix Pencils, Springer Berlin Heidelberg, 286–293, doi:10.1007/BFb0062108, 1983.

Yu, H., Kaufman, Y. J., Chin, M., Feingold, G., Remer, L. A., Anderson, T. L., Balkanski, Y., Bellouin, N., Boucher, O., Christopher, S., DeCola, P., Kahn, R., Koch, D., Loeb, N., Reddy, M. S., Schulz, M., Takemura, T., and Zhou, M.: A review of measurement-based assessments of the aerosol direct radiative effect and forcing, Atmos. Chem. Phys., 6, 613–666, doi:10.5194/acp-6-613-2006, 2006.

Table 1. IMPROVE network sites and number of samples in each dataset described in Sect. 2.1.

Site ID	Location	Type	Calibration 2011	Test 2011	Test 2013	Test 2013 Addl
1	Mesa Verde	Rural	79	40	112	0
2	Olympic	Rural	79	40	118	0
3	Phoenix	Urban	61	31	120	0
3B	Phoenix	Urban	64	33	120	0
4	Proctor Maple R. F.	Rural	70	35	122	0
4B	Proctor Maple R. F.	Rural	0	0	121	0
5	Sac and Fox	Rural	35	18	0	0
6	St. Marks	Rural	68	35	115	0
7	Trapper Creek	Rural	70	36	121	0
8	Birmingham	Urban	0	0	0	115
9	Bliss SP	Rural	0	0	0	85
10	South Korea	Urban	0	0	0	79
11	Fresno	Urban	0	0	0	116
12	Gr. Smoky Mt. NP	Rural	0	0	0	117
13	Hoover	Rural	0	0	0	86
14	Okefenokee NWR	Rural	0	0	0	117
15	Puget Sound	Urban	0	0	0	121
16	Cape Romain NWR	Rural	0	0	0	113
17	Tallgrass	Rural	0	0	0	110
18	Yosemite	Rural	0	0	0	115
18B	Yosemite	Rural	0	0	0	116
	Total		526	269	949	1290

Table 2. MDL and precision for FT-IR OC and TOR OC.

	TOR OC	FT-IR OC Calibration 2011	FT-IR OC Test 2011	FT-IR OC Test 2013	FT-IR OC Test 2013 Addl
MDL (µg m^{-3})[a]	0.05	013	0.16	0.10	0.10
% below MDL	1.5	2.5	6.0	3.3	0.6
Precision (µg m^{-3})[a]	0.14	0.08	0.13	0.07	0.14
Mean blank (µg)	NR[b]	-0.18 ± 1.4	-0.41 ± 1.8	1.78 ± 1.1	1.78 ± 1.1

[a] Concentration units of µg m^{-3} for MDL and precision are based on the IMPROVE volume of 32.8 m^3. [b] Not reported.

Table 3. FT-IR OC, anticipation of the prediction error per sample. Percentage of ambient samples falling in the FN (false negative), FP (false positive), TN (true negative) and TP (true positive) sectors. Performance metrics of the models that exclude predictions of samples that fall in the FP and TP sectors.

					Performance metrics			
Data set	FN	FP	TN	TP	Bias (μg m^{-3})*	Error (μg m^{-3})*	Norm. Error	R^2
Test 2011	0 %	0 %	99.6 %	0.4 %	0.01	0.08	11 %	0.96
Test 2013	0.2 %	0.2 %	99.5 %	0.1 %	0.09	0.12	18 %	0.96
Test 2013 Addl	1.7 %	0.6 %	94.5 %	3.2 %	0.05	0.12	14 %	0.92

* Concentration units of μg m^{-3} for Bias and Error are based on the IMPROVE volume of 32.8 m^3.

Table 4. MDL and precision for FT-IR EC and TOR EC.

	TOR EC	FT-IR EC Calibration 2011	FT-IR EC Test 2011	FT-IR EC Test 2013	FT-IR EC Test 2013 Addl
MDL (µg m^{-3})[a]	0.01[b]	0.02	0.02	0.01	0.01
% below MDL	3	0	1	0	0
Precision (µg m^{-3})[a]	0.11	0.04	0.04	0.03	0.04
Mean blank (µg)	NR[c]	0.02 ± 0.17	0.06 ± 0.17	0.06 ± 0.14	0.06 ± 0.14

[a] Concentration units of µg m^{-3} for MDL and precision are based on the IMPROVE volume of 32.8 m^3. [b] Value reported for network (0.44 µg) in concentration units. [c] Not reported.

Table 5. FT-IR EC, anticipation of the prediction error per sample. Percentage of ambient samples falling in the FN (false negative), FP (false positive), TN (true negative) and TP (true positive) sectors. Performance metrics of the models that exclude predictions of samples that fall in the FP and TP sectors.

					Performance metrics			
Data set	FN	FP	TN	TP	Bias (µg m^{-3})*	Error (µg m^{-3})*	Norm. Error	R^2
Test 2011	0 %	0 %	100 %	0 %	0	0.02	20 %	0.97
Test 2013	0.3 %	0.2 %	99.5 %	0 %	0	0.03	24 %	0.96
Test 2013 Addl	0.5 %	2.3 %	96.2 %	1 %	0.02	0.04	24 %	0.92

* Concentration units of µg m^{-3} for Bias and Error are based on the IMPROVE volume of 32.8 m^3.

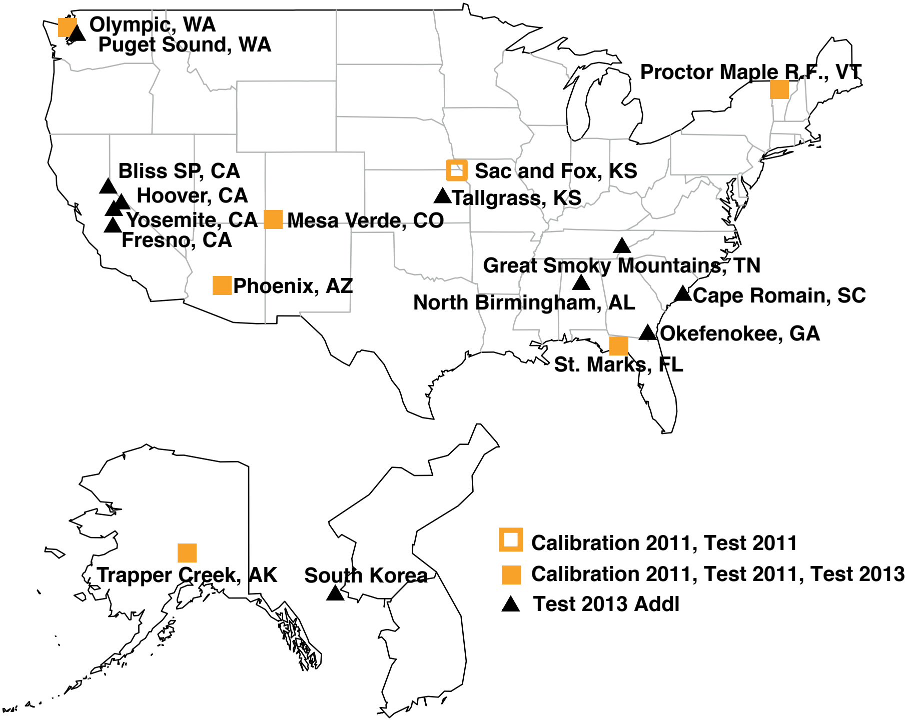

Figure 1. IMPROVE network sites.

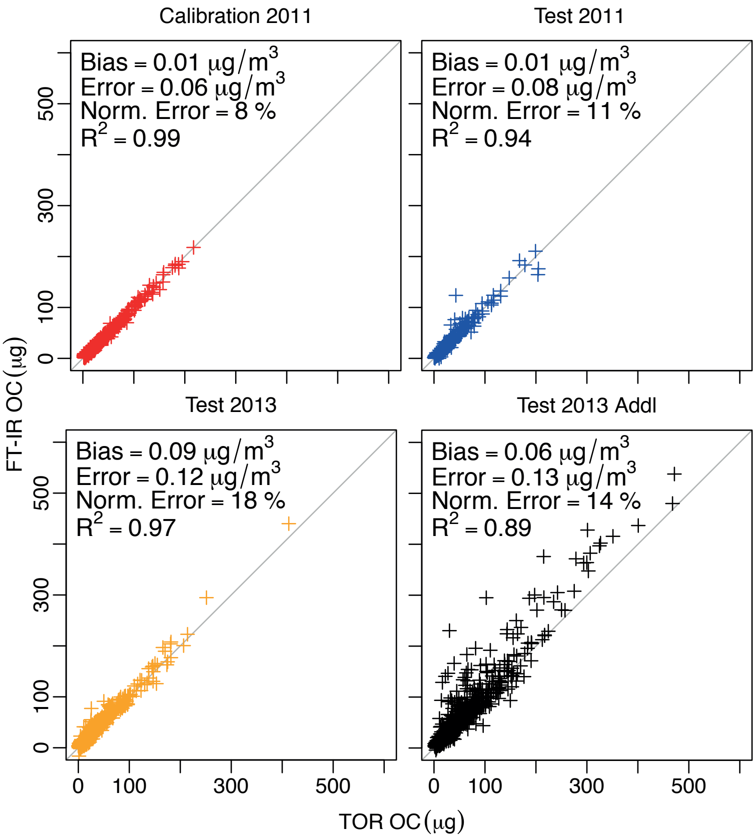

Figure 2. Scatterplots and performance metrics between FT-IR OC and TOR OC for the four datasets described in Sect. 2.1. Concentration units of μg m^{-3} for bias and error are based on the IMPROVE nominal volume of 32.8 m^3.

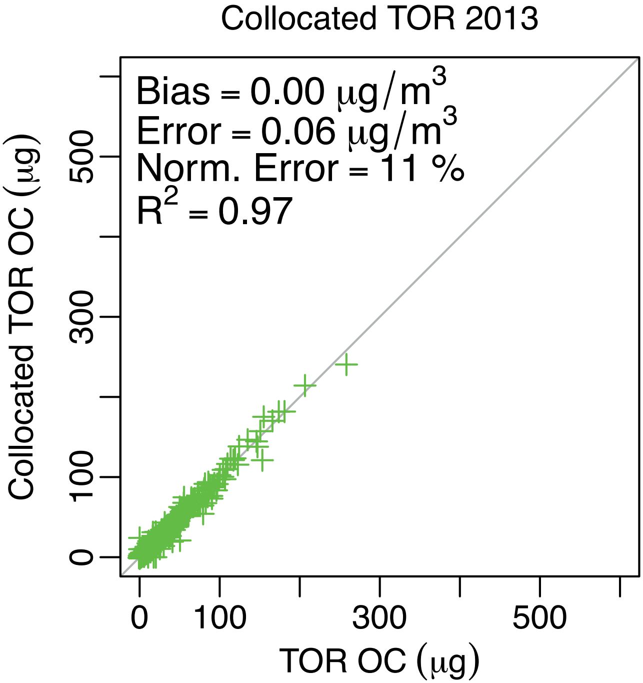

Figure 3. Scatterplot and performance metrics between between the measurements of the collocated TOR OC. Concentration units of $\mu g\, m^{-3}$ for bias and error are based on the IMPROVE nominal volume of $32.8\, m^3$.

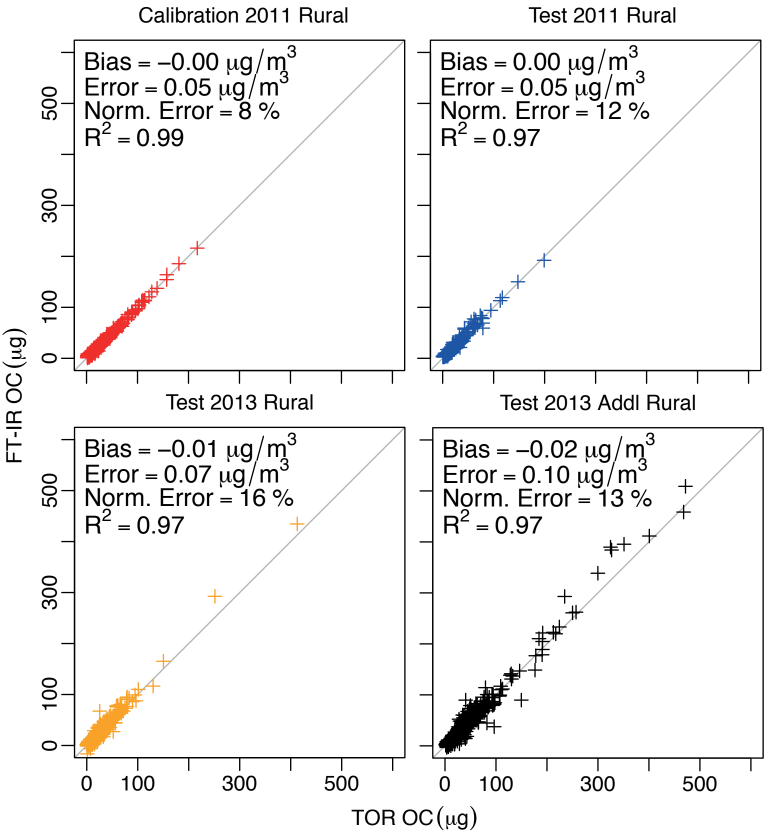

Figure 4. Scatterplots and performance metrics between FT-IR OC and TOR OC for the rural sites of the four datasets described in Sect. 2.1. Concentration units of µg m^{-3} for bias and error are based on the IMPROVE nominal volume of 32.8 m^3.

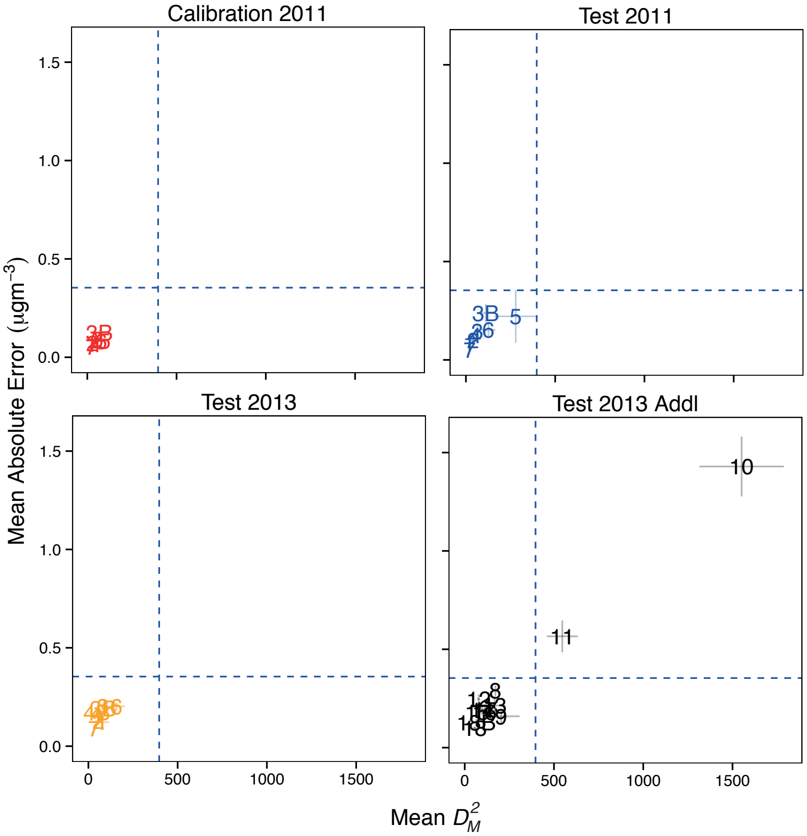

Figure 5. FT-IR OC. Anticipation of the prediction error. Mean squared Mahalanobis distance (between the scores of each dataset described in Sect. 2.1 and the centroid of the calibration dataset) against mean absolute error (between TOR OC and FT-IR OC). The measurements are are aggregated per site. Each site is denoted with the Site ID used in Table 1.

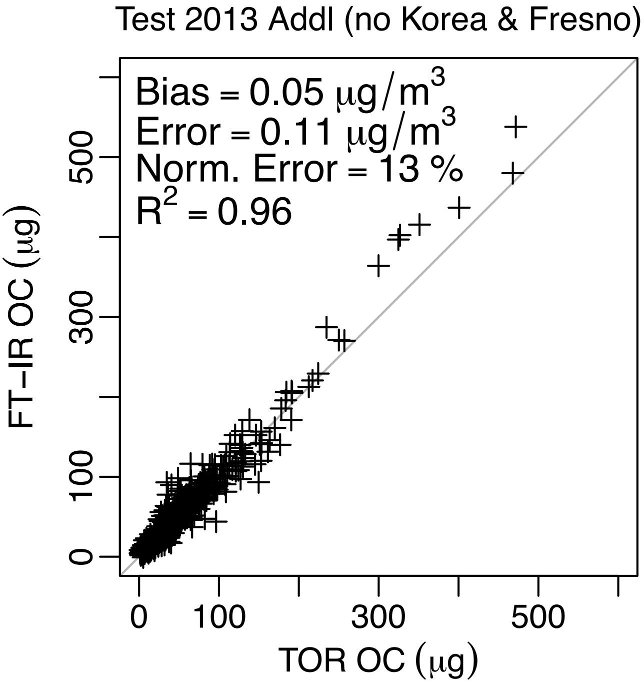

Figure 6. Scatterplot and performance metrics between FT-IR OC and TOR OC for the Test 2013 AddI datasets (Sect. 2.1) without the ambient samples of collected at Fresno and Korean sites. Concentration units of $\mu g\,m^{-3}$ for bias and error are based on the IMPROVE nominal volume of $32.8\,m^3$.

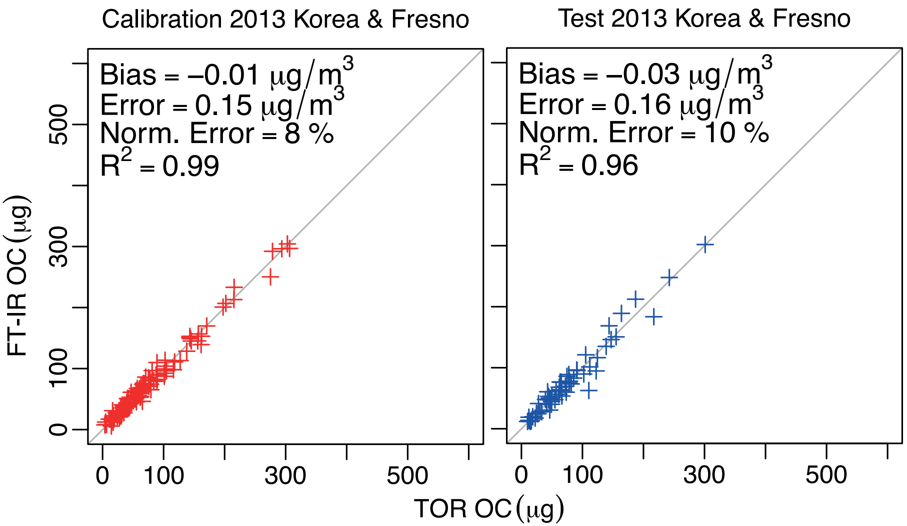

Figure 7. Scatterplot and performance metrics between FT-IR OC and TOR OC of the Korea and Fresno sites (Site ID 10 and 11 respectively). A new calibration set, based on two-thirds of the ambient samples collected at these two sites in 2013 is used for a dedicated model. Concentration units of μg m^{-3} for bias and error are based on the IMPROVE nominal volume of 32.8 m^3.

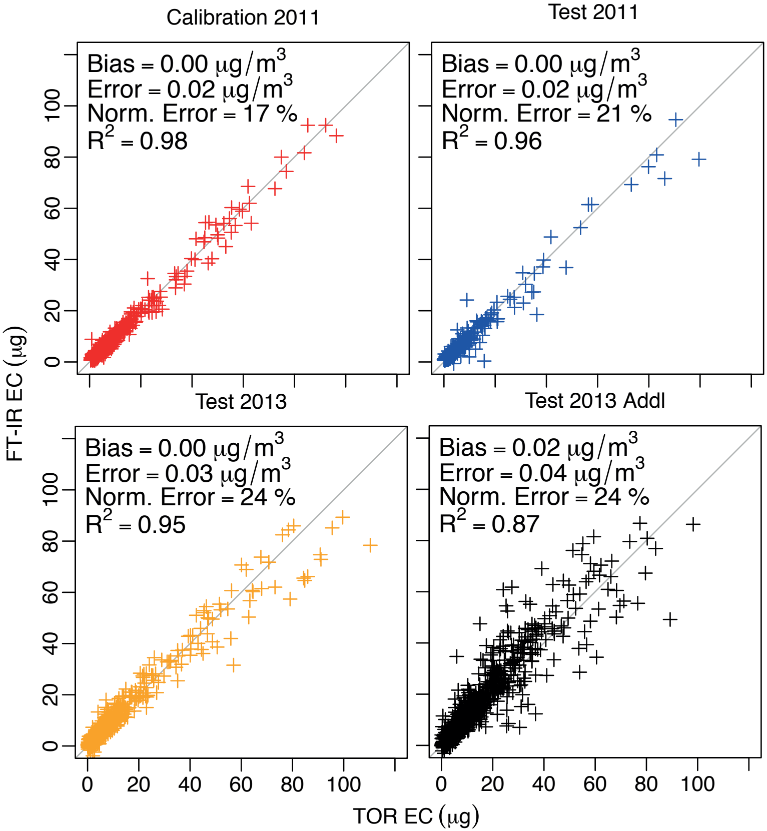

Figure 8. Scatterplots and performance metrics between FT-IR EC and TOR EC for the four datasets described in Sect. 2.1. Concentration units of µg m^{-3} for bias and error are based on the IMPROVE nominal volume of 32.8 m^{3}.

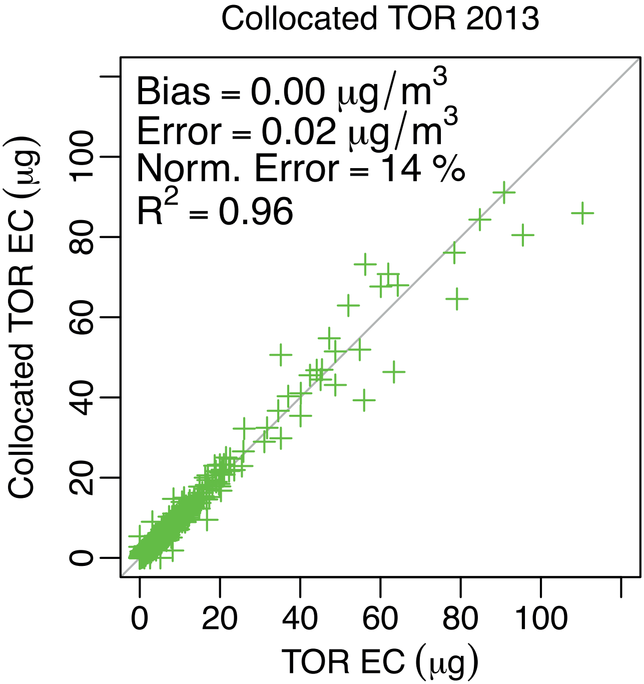

Figure 9. Scatterplot and performance metrics between between the measurements of the collocated TOR EC. Concentration units of μg m^{-3} for bias and error are based on the IMPROVE nominal volume of 32.8 m^3.

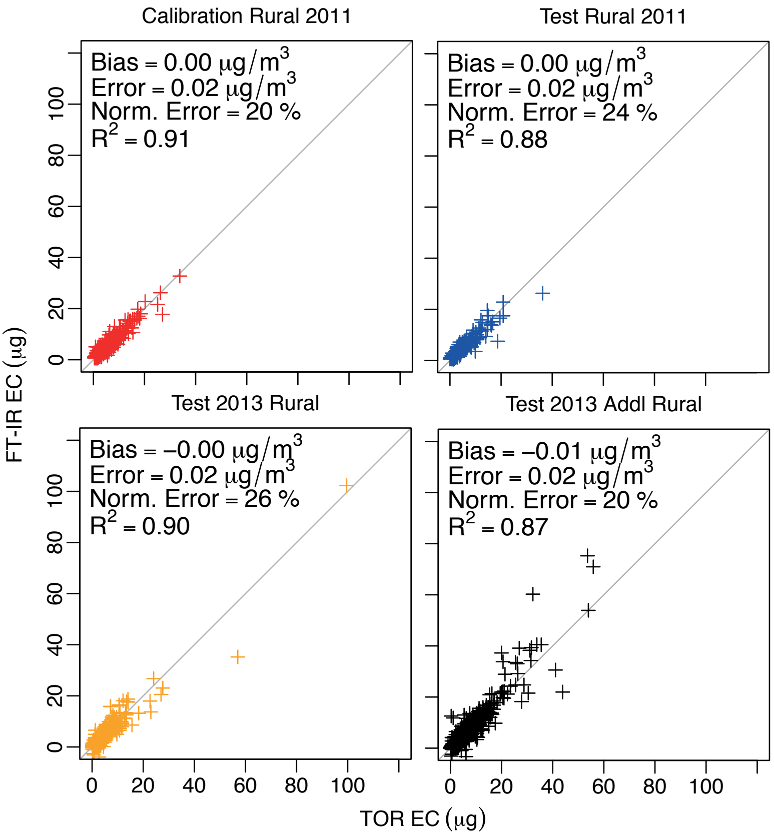

Figure 10. Scatterplots and performance metrics between FT-IR EC and TOR EC for the rural sites of the four datasets described in Sect. 2.1. Concentration units of µg m^{-3} for bias and error are based on the IMPROVE nominal volume of 32.8 m^3.

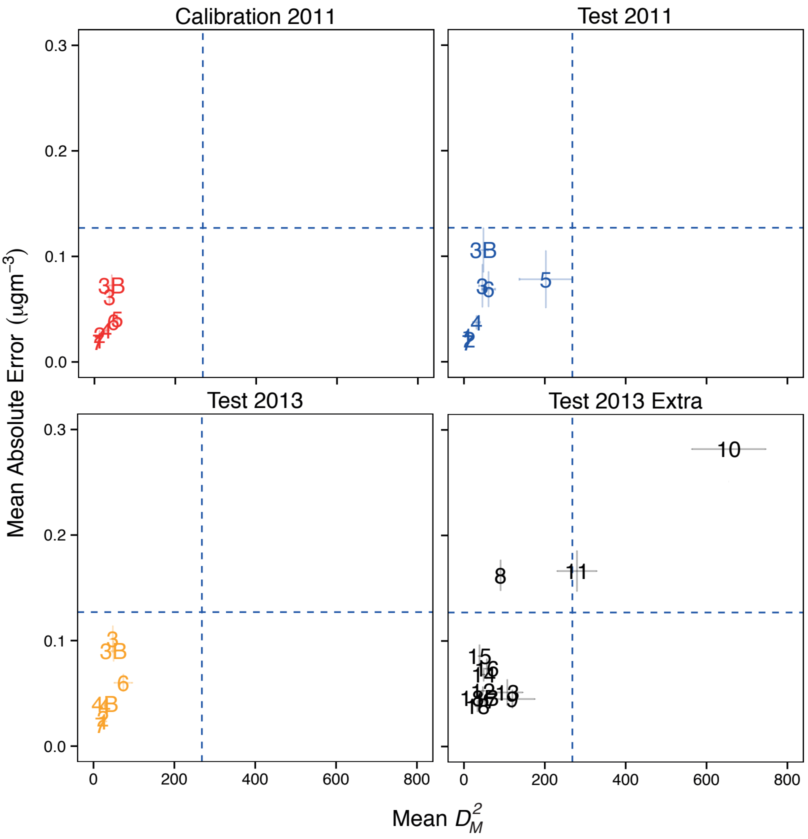

Figure 11. FT-IR EC. Anticipation of the prediction error. Mean squared Mahalanobis distance (between the scores of each dataset described in Sect. 2.1 and the centroid of the calibration dataset) against mean absolute error (between TOR OC and FT-IR OC). The measurements are are aggregated per site. Each site is denoted with the Site ID used in Table 1.

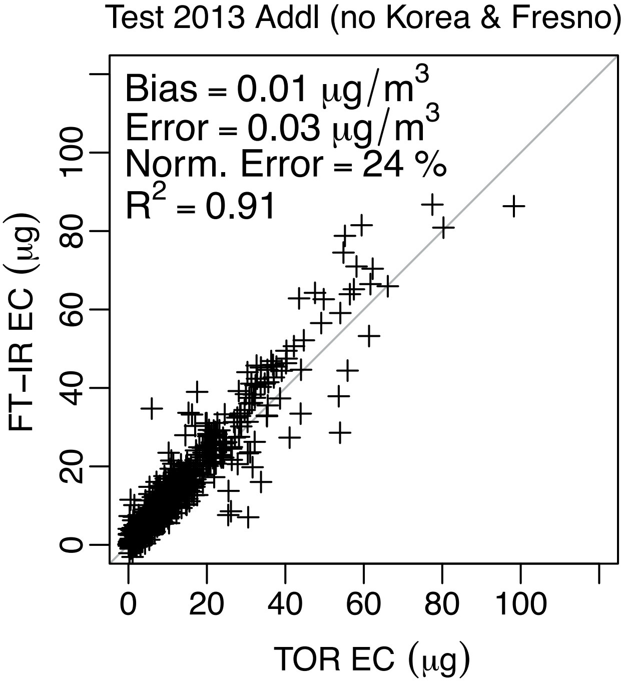

Figure 12. Scatterplot and performance metrics between FT-IR OC and TOR OC for the Test 2013 Addl datasets (Sect. 2.1) without the ambient samples of collected at Fresno and Korean sites. Concentration units of μg m^{-3} for bias and error are based on the IMPROVE nominal volume of 32.8 m^3.

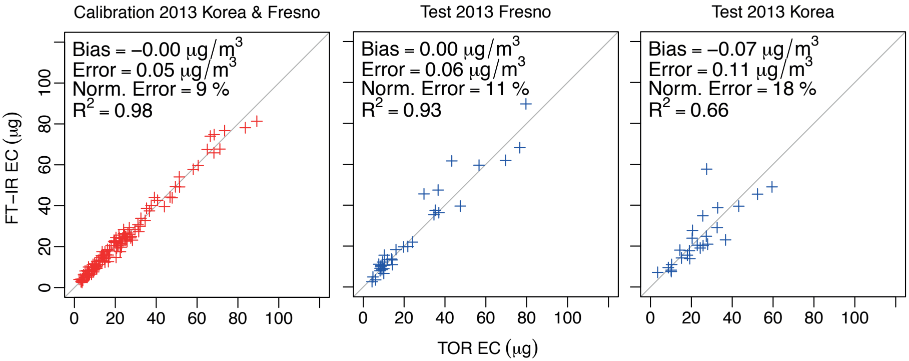

Figure 13. Scatterplot and performance metrics between FT-IR EC and TOR EC of the Korea and Fresno sites (Site ID 10 and 11 respectively). A new calibration set, based on two-thirds of the ambient samples collected at these two sites in 2013 is used for a dedicated model. Concentration units of µg m^{-3} for bias and error are based on the IMPROVE nominal volume of 32.8 m^3.