Point-by-point response to review comments on manuscript acp-2015-997
"Determination of Primary combustion source organic carbon-to-elemental carbon (OC/EC) ratio using ambient OC and EC measurements: Secondary OC-EC correlation minimization method"

By Cheng Wu and Jian Zhen Yu

We thank the two anonymous reviewers for their constructive comments. Our point-by-point responses to the review comments are listed below. Changes to the manuscript are marked in blue in the revised manuscript. The marked manuscript is submitted together with this response document.

Anonymous Referee #1

Dear Editor, this MS presents a statistical assessment of an alternative method to quantify secondary organic carbon (SOC) in ambient air samples. This method is an alternative to the classic EC tracer method. It is a useful assessment of an alternative method which seems to perform rather well, and therefore merits publication. Reading is somewhat complicated due to the frequent use of abbreviations (eg, fSOC), though. A more fluent writing style would help the reader.

Author's Response: We add a table (also shown below) in the revised main text to help readers to have a quick check of abbreviations used in the paper. We believe this would be more reader-friendly than looking for definitions that scattered in the main text. Please see below for point-by-point response to reviewers' comments.

Table 1. Acronyms and Abbreviations

Abbreviation	Definition
EC	elemental carbon
EC_1, EC_2	EC from source 1 and source 2 in the two-source scenario
f_{EC1}	fraction of EC from source 1 to the total EC
f_{SOC}	ratio of SOC to OC
MRS	minimum R squared method
MRS'	A variant of MRS that use EC from individual sources as input
MT	Mersenne twister pseudorandom number generator
n	sample size in MT data generation
OC	organic carbon
OC/EC	OC to EC ratio
$(OC/EC)_{pri}$	primary OC/EC
$OC/EC_{10\%}$	OC/EC at 10% percentile
OC/EC_{min}	minimum OC/EC
$OC_{non\text{-}comb}$	OC from non-combustion sources
PDF	probability density function of a distribution
POC	primary organic carbon
ROA	ratio of averages
RSD	relative standard deviation
RSD_{EC}	RSD of EC
RSD_{POC}	RSD of POC
RSD_{SOC}	RSD of SOC
SOC	secondary organic carbon
SOC_{svP}	SOC formed from semi-volatile POC
Υ_pri	ratio of the $(OC/EC)_{pri}$ of source 2 to source 1
ε_{EC} , ε_{OC}	measurement uncertainty of EC and OC
Υ_{unc}	relative measurement uncertainty
$\gamma\__{RSD}$	the ratio between the RSD values of $(OC/EC)_{pri}$ and EC

Some specific comments:

- line 75: I believe Pio et al propose yet another method, using a subset of samples with 5% lowest ratios and discarding the 3 lowest... I don't have the exact reference right now, but please add.

Author's Response: Suggestion taken. The reference (Pio et al., 2011) is now added in the main text.

> Lines 74-78
> "Combinations of the fixed percentile and the minimum $(OC/EC)_{pri}$ approaches were also used in order to accommodate different sample sizes available. For example, Pio et al. (2011) suggested using the lowest 5% subset to obtain the $(OC/EC)_{pri}$, and if the sample size of 5% subset is less than three, the lowest three data points are used to determine $(OC/EC)_{pri}$."

Reference

Pio, C., Cerqueira, M., Harrison, R. M., Nunes, T., Mirante, F., Alves, C., Oliveira, C., de la Campa, A. S., Artinano, B., and Matos, M.: OC/EC ratio observations in Europe: Re-thinking the approach for apportionment between primary and secondary organic carbon, Atmos Environ, 45, 6121-6132, DOI 10.1016/j.atmosenv.2011.08.045, 2011.

- line 90: any reason why the Millet method was overlooked?

Author's Response: One reason is that Millet's original paper focused on VOCs, and the MRS approach was used to calculate primary ratio of VOCs/EC to differentiate primary and secondary VOCs. A second reason we believe is a lack of evaluation work for this method. As a result, the approach initially proposed by Millet et al did not draw much attention from the OC/EC measurement community.

- line 211: please elaborate on why the OCEC10% method provides worse results

Author's Response: Based on the observational data we have, the ambient conditions most likely falls into the scenario between scenario A and B (Figure 3). As such, $OC/EC_{10\%}$ is further away from the true OC/EC_{pri} than OC/EC_{min}, resulting larger bias.

- line 226: I don't understand the different behavior of the $OCEC_{10\%}$ amend the $OCEC_{min}$ methods, given that they are both subsets of the total dataset with specific characteristics of representing 1% and 10%. Why is their behavior different?

Author's Response: Change of f_{SOC} not only changes the position of OC/EC distribution relative to OC/EC_{pri} distribution, but can also alter the width of OC/EC distribution. Because the subset methods rely on percentile of OC/EC, once the OC/EC distribution is widened, the relative position between OC/EC_{min} and $OC/EC_{10\%}$ is also changed and this results in a non-linear response in SOC differences,

- section uncertainty: with some analytical methods (e.g., TOT) the uncertainty is mostly constant (0,1-0,2 micrograms/cm2), please discuss how this would affect the results in this section.

Author's Response: Under the scenario of constant absolute uncertainty, the performance of MRS (Figure R1, 0.2 μg m^{-3}) is similar to that assuming a fix proportional measurement uncertainty (Figure R2, 10% measurement uncertainty). Both Figures R1 and R2 are included in the revised main text as Figure 8.

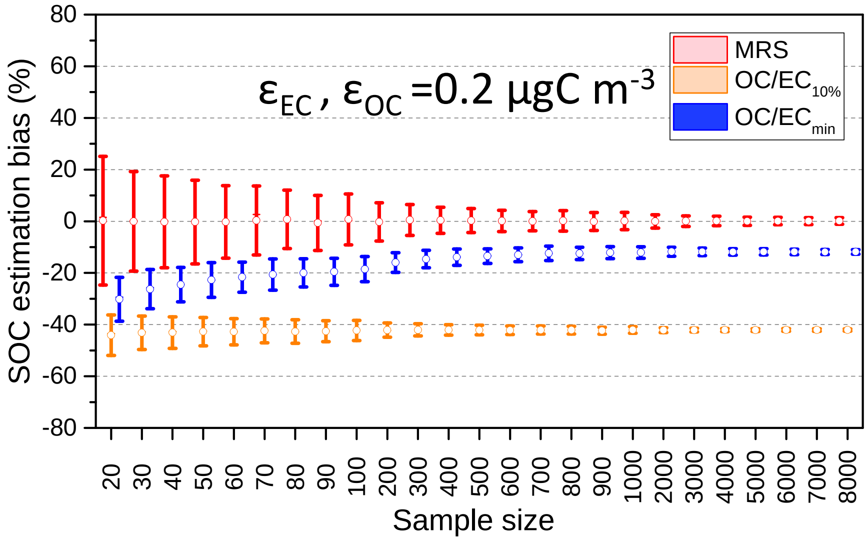

Figure R1. SOC estimation bias as a function of sample size, assuming fixed absolute measurement uncertainty for OC and EC (0.2 μgC m^{-3}). For each sample size, 500 repeat runs were conducted. The circles represent mean of 500 repeat runs, the whiskers represent one standard deviation. Parameters used for testing: Repeat runs = 500; N = 20~8000; EC = 8±4 μgC m^{-3}; $(OC/EC)_{pri}$ = 0.5; POC = 4 ±2 μgC m^{-3}, f_{SOC} =40%, and SOC = 2.67±1.33 μgC m^{-3}.

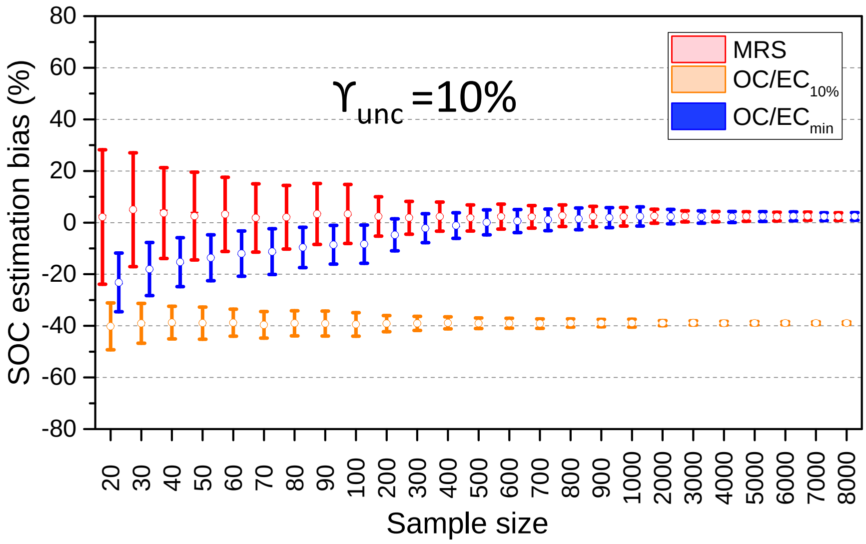

Figure R2. SOC estimation bias as a function of sample size, assuming a fixed relative measurement uncertainty of 10% for OC and EC. For each sample size, 500 repeat runs were conducted. The open circle represents the mean of 500 repeat runs, and the whisker represents one standard deviation. Parameters used for testing: Repeat runs = 500; N = 8000; EC = 8±4 μgC m^{-3}; $(OC/EC)_{pri}$ = 0.5; POC = 4 ±2 μgC m^{-3}, f_{SOC} =40%, and SOC = 2.67±1.33 μgC m^{-3}.

- line 317, please clarify what the authors mean by "the irrelevance of EC and SOC", it is unclear to me

Author's Response: We now rephrased as “the independence of EC and SOC”, by which we mean that SOC and EC come from uncorrelated sources.

Anonymous Referee #2

Generally Comments

Typically the EC tracer method, when used in estimating the secondary organic carbon (SOC), relies on three conditions– 1) the relatively constant $(OC/EC)_{pri}$ over the period of study; 2) the random nature of SOC formation relative to EC; and 3) a subset of dataset without significant SOC contributions. The $OC/EC_{10\%}$ or OC/EC_{min} essentially utilize the subset in Condition #3 to derive the $(OC/EC)_{pri}$ if it does have an unique value. Any deviations from the conditions as well as measurement uncertainties will lead to bias in determining $(OC/EC)_{pri}$. In some environments where SOC dominates, the third condition is generally impossible to be met. This study, through an extensive test, shows that the third condition is not necessary in calculating $(OC/EC)_{pri}$, if an algorithm, i.e., minimum R^2 (MRS), is used looking for $(OC/EC)_{pri}$ that yields SOC least correlated with EC. Without further examinations, the reviewer thinks that MRS is probably mathematically rigorous for any datasets satisfying the first two conditions and, additionally, with sufficient size and accuracy. It can perform better than $OC/EC_{10\%}$ or OC/EC_{min} most of the time because Condition 3 is fortuitous, as described by the authors.

While the reviewer agrees that MRS should be used instead of $OC/EC_{10\%}$ or OC/EC_{min} in calculating SOC, particularly for a large dataset which can support meaningful correlation analysis, MRS does not solve fundamental problems in the EC tracer method. The $(OC/EC)_{pri}$ is by no means constant, as it varies with source contributions from day to day and season to season. SOC is likely correlated with EC because in urban areas many SOC precursors originate from the same combustion sources as EC. This paper demonstrates that when Conditions 1 and 2 are in doubt, MRS produces erroneous results. MRS results are also sensitive to measurement uncertainty that impacts the correlation coefficients. These limitations, however, are not emphasized adequately in the abstract, which sounds almost like MRS has tackled all these issues. These issues, still, can only be solved by using multivariate or chemical mass balance analysis with additional markers.

Author's Response: Thanks for the very insightful comments. We agree that $(OC/EC)_{pri}$ varied from day to day and season to season in reality and this limitation is intrinsic in the EC tracer method regardless different approaches in implementing the EC tracer method, unless it is applied in a time frame small enough that variations of $(OC/EC)_{pri}$ is almost negligible. Limits posed by the nature of ambient ECOC data are inherent to the EC tracer method and common to all the variants of the EC tracer method. This study focuses on evaluating different $(OC/EC)_{pri}$ determination approaches within the EC tracer method, with the aim to identify the best approach in applying the EC tracer method. We have revised the wording in the abstract and in main text to emphasize the limitations of the EC tracer method and the MRS approach. Please see below the specific revisions in our point-by-point response to reviewers' comments.

Specific Comments

Abstract: Please describe the assumptions of MRS, datasets that are suitable for MRS analysis, and potential errors while in the same time shortening the abstract. Just saying MRS is better than $OC/EC_{10\%}$ or OC/EC_{min} is not meaningful because all the three could be very wrong in some cases.

Author's Response: We have made the following revisions in the abstract to clearly state the assumptions of MRS.

Line 28:

"The hypothetical $(OC/EC)_{pri}$ that generates the minimum R^2(SOC,EC) then represents the actual $(OC/EC)_{pri}$ ratio if variations of EC and SOC are independent and $(OC/EC)_{pri}$ is relatively constant in the study period."

Line 38-41:

"…MRS provides an unbiased SOC estimation when measurement uncertainty is small. MRS results are sensitive to the magnitude of measurement uncertainty but the bias would not exceed 23% if the uncertainty is controlled within 20%."

We also shortened slightly the abstract by condensing a few sentences and removing the following sentence (this background information is spelled out in the introduction section).

~~"The general concept embodied in the MRS method was initially proposed by Miller et al (2005), but has not been evaluated for accuracy or utility since its debut."~~

Line 97-102: While using simulated data is insightful, it offers no proof. The authors may explore if there is a true "proof" from mathematical or statistical derivations that MRS will yield true $(OC/EC)_{pri}$ if SOC is indeed random and the dataset is big enough. This may also answer the question- how big is big? MRS does not seem suitable for a dataset with only dozens of points.

Author's Response:

We agree that the simulated data alone does not offer proof, as there is no guarantee that the simulated data capture all the essential features of real-world data. In response to this comment, we conducted a series of sensitivity tests to evaluate the SOC estimation dependency on sample size, which was varied from 20 to 8000. For each sample size, 500 repeat runs were tested, assuming a single value OC/EC_{pri} with a measurement uncertainty of 10%. The results are in Fig. R2, showing the average and the standard deviation for each sample size. The standard variation of SOC bias by MRS decreases with increased sample size while the mean of SOC bias remains a constant small value (2%). The standard variation of SOC bias is ~±30% at the lowest tested sample size (n = 20), and decreases to less than 15% at n =60 (the sample size of one-year sampling from an every-six-day sampling program) and to less than 10% at n = 200. Other scenarios considering OC/EC_{pri} with a distribution and different f_{SOC} are discussed in SI. Figure R2 will be included in the main text.

A new section (as shown below) is added to the manuscript to address the sample size question.

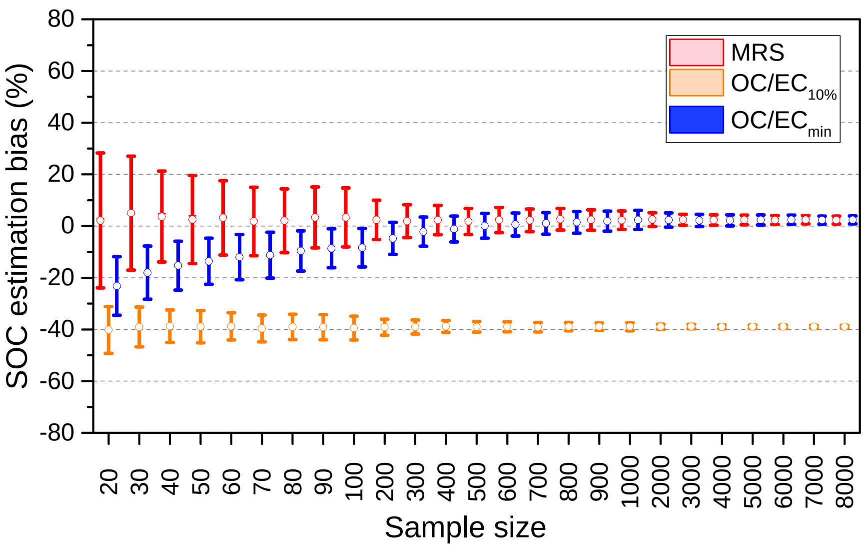

Figure R2. SOC estimation bias as a function of sample size, assuming a fixed relative measurement uncertainty of 10% for OC and EC. For each sample size, 500 repeat runs were conducted. The open circle represents the mean of 500 repeat runs, and the whisker represents one standard deviation. Parameters used for testing: Repeat runs = 500; N = 8000; EC = 8±4 μgC m^{-3}; $(OC/EC)_{pri}$ = 0.5; POC = 1 ±0.5 μgC m^{-3}, f_{SOC} =40%, and SOC = 0.67±0.34 μgC m^{-3}.

Lines 331-356:

"**2.4 Impact of sample size**

MRS rely on correlations of input variables and it is expected that MRS performance is sensitive to the sample size of input dataset. This section examines the sensitivity on sample size by the three $(OC/EC)_{pri}$ representations and aims to provide suggestions for an appropriate sample size when applying MRS on ambient OCEC data. Sample sizes ranging from 20 ~ 8000 are tested and for each sample size 500 repeat runs are conducted to obtain statistically significant results. Both Case A (i.e., a constant relative uncertainty of 10%) and Case B (i.e., a constant absolute uncertainty of ±0.2 μgC m^{-3} for both OC and EC) are considered. The measurement uncertainties in case B are generated separately by MT following a uniform distribution within the range of ±0.2 μgC m^{-3}. The measurement uncertainties of POC and SOC are then back-calculated following the uncertainty propagation formula (Harris, 2010) and assuming the ratio of $\varepsilon_{POC}/\varepsilon_{SOC}$ is the same as POC/SOC ratio (controlled by f_{SOC}).

The mean SOC bias by MRS is very small (<3%) for all sample sizes while the standard deviation of SOC bias decreases with sample size (Figure 8). The standard deviation of SOC bias is ~±30% at the lowest test sample size (n = 20), and decreases to less than ±15% at n = 60 (the sample size of one-year sampling from an every-six-day sampling program) and to less than ±10% at n = 200. Similar patterns are observed between Case A (Figure 8a) and Case B (Figure 8b) for MRS and $OC/EC_{10\%}$. For OC/EC_{min}, a larger bias is observed in Case B than Case A for all sample sizes, as SOC bias by OC/EC_{min} is more sensitive to measurement uncertainty in the range of 0~10% as shown in Figure 7b. The standard deviation of SOC bias by OC/EC_{min} and $OC/EC_{10\%}$ both decreases with sample size as shown in Figure 8. The mean SOC bias of OC/EC_{min} decrease with increased sample size while $OC/EC_{10\%}$ is insensitive to sample size. The sample size dependency of all three $(OC/EC)_{pri}$ representations is not sensitive to f_{SOC} as shown in Figure S16. Other scenarios considering $(OC/EC)_{pri}$ with a distribution and different f_{SOC} are discussed in SI."

Line 116-118: How good are the K-S statistics? In other words, how well did the pseudorandom number generator reproduce the statistics in the original dataset?

Author's Response: The K-S statistics for ambient measured data are shown in Figures S1-S4 (This information is now also mentioned in the main text). In Igor Pro's Kolmogorov–Smirnov test, D represents the K-S statistic, C represent critical value. If D<C, the samples follow the corresponding distribution (e.g., normal or log-normal distribution). The majority of the data can pass the K-S test for log-normal distribution and some exhibit a bimodal distribution. For the performance of the MT pseudorandom number generator, we conduct a series of K-S tests on the generated data for 5000 runs, which show 94.4% data having D small than C (Fig. R3). Hence, we believe the pseudorandom number generator could produce the data following preset characteristics. Figure R3 is added to the SI and referred to in the main text.

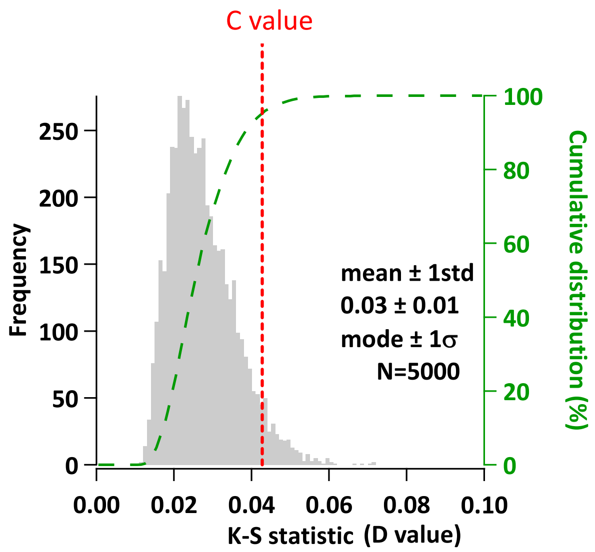

Figure R3 Performance of the MT pseudorandom number generator evaluated by K-S test. The histogram in grey represents D statistic value in K-S test and the red dashed-line represents C. The dash line in green represents cumulative distribution of D. Data with D<C, i.e., data that strictly follow the log-normal distribution, account for 94.4% in 5000 runs.

The below text is added to the manuscript to describe whether the pseudorandom number generate reproduce the statistics in the original dataset.

> Lines 142-145:
> "For the verification of the log-normality of MT generated data, a series of K-S tests on the generated data for 5000 runs are conducted. As shown in Figure S6, 94.4% of runs pass the K-S test. Hence the performance of MT can satisfy the log-normal distributed data generation requirement in this study".

Line 126: Eqs. (4)-(5) do not work for all datasets. They are probably asymptotes when datasets are large enough in size.

Author's Response: We agree that they do not necessarily work for all datasets. The reason for translating mean and standard deviations into μ and σ is that the MT pseudorandom number generator in Igor Pro only accepts μ and σ as input parameters, while mean and standard deviations are the parameters that can be obtained from ambient measurements.

Line 136: Mention here that the case with combustion-related SOC is discussed later.

Author's Response: Suggestion taken. The text below is included in the revised manuscript:

> Line 140:
> "The case with combustion-related SOC is briefly discussed in section 3."

Line 151-152: The results of log-normally distributed $(OE/CC)_{pri}$ should be summarized in the text if possible.

Author's Response: Suggestion taken. The below text is added to the section 2.2.1:

> Lines 219-225:
> "For the representation of $(OC/EC)_{pri}$ in the simulated data as lognormally distributed data, analysis is also performed to evaluate SOC estimation bias as a function of RSD_{EC}, RSD_{SOC}, and f_{SOC}. Table

S2 summarizes the results obtained with adopting most probable ambient conditions (i.e., RSD_{EC}: 50-100%, f_{SOC}: 40-60%). SOC bias by MRS is within 4% when measurement uncertainty is ignored. In comparison, SOC bias by OC/EC_{min} is more sensitive to assumption of log-normally distributed $(OC/EC)_{pri}$ than single value $(OC/EC)_{pri}$, including the dependency on RSD_{EC} and RSD_{SOC} with varied $f_{SOC..}$"

Line 220-222: It is not clear if f_{EC1} was varied from sample to sample in a single test or only varied from test to test. If the former, how could you make sure EC1 and EC2 are highly correlated?

Author's Response: f_{EC1} was varied from test to test. The text is now clarified as below:

Lines 229-231:
"By varying f_{EC1} (proportion of source 1 EC to total EC) from test to test, the effect of different mixing ratios of the two sources can be examined."

Line 284-286: Since POC and SOC are not directly measured, what is the meaning to simulate their measurement uncertainty?

Author's Response: Once OC and EC data are considered to have measurement uncertainty, the derived quantities POC and SOC (using Eq (1) and Eq (2)) consequently also have associated uncertainty, which can be calculated following uncertainty propagation principle. For the evaluation of SOC estimation, SOC calculated from the EC tracer method needs to be compared with "true SOC plus associated uncertainty". That's the reason why we calculated the uncertainties of POC and SOC

Line 384: How were the six subsets selected?

Author's Response: With a given one-year data set, there are six possible extractions of daily data sets corresponding to the assumed every-six-day sampling schedule, i.e., set 1:{Day 1, 7, 13,..}, set 2: {Day 2, 8, 14,..}, set 3: {Day 3, 9, 15,..}, etc. The text below is added to clarify this point:

Lines 368-371:
"The one-year data yields six subsets of daily samples, corresponding to six possible schedules of sampling days with the every-six-day sampling frequency. The MRS calculation produces the OC/EC_{pri} in the range of 2.37 – 2.75…"

Line 360-362: Emphasize that this only happens when measurement uncertainties are small.

Author's Response: Suggestion taken. This sentence is revised as below:

Lines 408-413:
"In the scenarios of a single primary source and two well-correlated primary combustion sources, SOC estimates by MRS are unbiased while OC/EC_{min} and $OC/EC_{10\%}$ consistently underestimate SOC when measurement uncertainty is neglected. When measurement uncertainty is considered, all three approaches produce biased SOC estimates, with MRS producing the smallest bias. The bias by MRS is less than 25% when measurement uncertainty is within 20% and f_{SOC} is not lower than 20%."

Determination of Primary combustion source organic carbon-to-elemental carbon (OC/EC) ratio using ambient OC and EC measurements: Secondary OC-EC correlation minimization method

Cheng Wu[1], Jian Zhen Yu[1, 2, 3]

[1] Division of Environment, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China

[2] Atmospheric Research Centre, Fok Ying Tung Graduate School, Hong Kong University of Science and Technology, Nansha, China

[3] Department of Chemistry, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China

Correspondence to: Jian Zhen Yu (jian.yu@ust.hk)

Abstract

Elemental carbon (EC) has been widely used as a tracer to track the portion of co-emitted primary organic carbon (OC) and, by extension, to estimate secondary OC (SOC) from ambient observations of EC and OC. Key to this EC tracer method is to determine an appropriate OC/EC ratio that represents primary combustion emission sources (i.e., $(OC/EC)_{pri}$) at the observation site. The conventional approaches include regressing OC against EC within a fixed percentile of the lowest (OC/EC) ratio data (usually 5-20%) or relying on a subset of sampling days with low photochemical activity and dominated by local emissions. The drawback of these approaches is rooted in its empirical nature, i.e., a lack of clear quantitative criteria in the selection of data subsets for the $(OC/EC)_{pri}$ determination. We examine here a method that derives $(OC/EC)_{pri}$ through calculating a hypothetical set of $(OC/EC)_{pri}$ and SOC followed by seeking the minimum of the coefficient of correlation (R^2) between SOC and EC. The hypothetical $(OC/EC)_{pri}$ that generates the minimum R^2(SOC,EC) then represents the actual $(OC/EC)_{pri}$ ratio if variations of EC and SOC are independent and $(OC/EC)_{pri}$ is relatively constant in the study period. This Minimum R Squared (MRS) method has a clear quantitative criterion for the $(OC/EC)_{pri}$ calculation. This work uses numerically simulated data to evaluate the accuracy of SOC estimation by the MRS method

and to compare with two commonly used methods: minimum OC/EC (OC/EC_{min}) and OC/EC percentile ($OC/EC_{10\%}$). Log-normally distributed EC and OC concentrations with known proportion of SOC are numerically produced through a pseudorandom number generator. Three scenarios are considered, including a single primary source, two independent primary sources, and two correlated primary sources. The MRS method consistently yields the most accurate SOC estimation. Unbiased SOC estimation by OC/EC_{min} and $OC/EC_{10\%}$ only occurs when the left tail of OC/EC distribution is aligned with the peak of the $(OC/EC)_{pri}$ distribution, which is fortuitous rather than norm. In contrast, MRS provides an unbiased SOC estimation when measurement uncertainty is small. MRS results are sensitive to the magnitude of measurement uncertainty but the bias would not exceed 23% if the uncertainty is within 20%.

1 Introduction

Organic carbon (OC) and elemental carbon (EC) are among the major components of fine particular matter ($PM_{2.5}$) (Malm et al., 2004). EC is a product of carbon fuel-based combustion processes and is exclusively associated with primary emissions whereas OC can be from both direct emissions and be formed through secondary pathways. Differentiation between primary organic carbon (POC) and secondary organic carbon (SOC) is indispensable for probing atmospheric aging processes of organic aerosols and formulating effective emission control policies. However, direct SOC measurement is not yet feasible, as there lacks knowledge of its chemical composition at the molecular level. Due to its exclusive origin in primary combustion sources, EC was first proposed by Turpin and Huntzicker (1991) to serve as the tracer to track POC from primary combustion sources and, by extension, to estimate SOC as SOC is simply the difference between OC and POC. This EC tracer method only requires measurements of OC and EC. Due to its simplicity, the EC tracer method has been widely adopted in studies reporting ambient OC and EC measurements (e.g., Castro et al., 1999;Cao et al., 2004;Yu et al., 2004). If OC and EC concentrations are available and primary OC from non-combustion sources ($OC_{non\text{-}comb}$) is negligible, SOC can be estimated using EC as the tracer for combustion source POC (Turpin and Huntzicker, 1995):

$$POC = (OC/EC)_{pri} \times EC \tag{1}$$

$$SOC = OC_{total} - (OC/EC)_{pri} \times EC \tag{2}$$

where $(OC/EC)_{pri}$ is the OC/EC ratio in freshly emitted combustion aerosols, and OC_{total} and EC are available from ambient measurements. Abbreviations used in this study are summarized in Table 1.

The key step in the EC tracer method is to determine an appropriate OC/EC ratio that represents primary combustion emission sources (i.e., $(OC/EC)_{pri}$) at the observation site. Various approaches in deriving $(OC/EC)_{pri}$ reported in the literature are either based on emission inventory (Gray et al., 1986) or ambient observation data. Using ambient observation data, three approaches are the most common: 1) regressing measured OC vs. EC data from times of low photochemical activity and dominated by local emissions; 2) regressing measured OC vs. EC data on a fixed percentile of the lowest OC/EC ratio (usually 5-20%) data to represent samples dominated by primary emissions (Lim and Turpin, 2002;Lin et al., 2009;Pio et al., 2011); and 3) simply taking the minimum OC/EC ratio during the study period to approximate $(OC/EC)_{pri}$ (Castro et al., 1999). Combinations of the fixed percentile and the minimum $(OC/EC)_{pri}$ approaches were also used in order to accommodate different sample sizes available. For example, Pio et al. (2011) suggested using the lowest 5% subset to obtain the $(OC/EC)_{pri}$, and if the sample size of 5% subset is less than three, the lowest three data points are used to determine $(OC/EC)_{pri}$. These approaches have the drawback in that there is not a clear quantitative criterion in the data selection for the $(OC/EC)_{pri}$ determination. Millet et al. (2005) was the first to propose an algorithm that explores the inherent independency between pollutants from primary emissions (e.g., EC) and products of secondary formation processes (e.g., SOC) to derive the primary ratios (e.g., $(OC/EC)_{pri}$) for species with multiple source types. More specifically, for the determination of $(OC/EC)_{pri}$, the assumed $(OC/EC)_{pri}$ value is varied continuously. At each hypothetical $(OC/EC)_{pri}$, SOC is calculated for the data set and a correlation coefficient value (R^2) of EC vs. SOC (i.e., R^2(EC,SOC)) is generated. The series of R^2(EC,SOC) values are then plotted against the assumed $(OC/EC)_{pri}$ values. If variations of EC and SOC are independent, the assumed $(OC/EC)_{pri}$ corresponding to the minimum R^2(EC,SOC) would then represent the actual $(OC/EC)_{pri}$ ratio. Such an approach obviates the need for an arbitrary selection criterion, as the algorithm seeks the minimum point, which is unique to the dataset. However, this method has largely been overlooked, with only one study reporting its use (Hu et al., 2012) since its debut, which may be a result of a lack of evaluation of its method performance. Hereafter for the convenience of discussion, we call this method the minimum R squared (MRS) method, with a conceptual illustration of the MRS method shown in Figure 1. A computer program written in Igor Pro (WaveMetrics, Inc. Lake Oswego, OR, USA) is developed to feasible MRS calculation and it is available from https://sites.google.com/site/wuchengust.

With ambient OC and EC samples, the accuracy of estimated SOC by different $(OC/EC)_{pri}$ methods is difficult to evaluate due to the lack of a direct SOC measurement. The objective of this study is to investigate, through numerical simulations, the bias of SOC estimates by three different implementations of the EC tracer method. Hypothetic EC, OC, and $(OC/EC)_{pri}$ datasets with known break-down of POC and SOC values are numerically synthesized, then SOC is estimated and compared with the "true" SOC as defined by the synthetic datasets. As such, bias of SOC estimates using the various implementations of the EC tracer method can be quantified.

2 Evaluation of the Minimum R Squared Method

2.1 Data generation

We first examine ambient OC and EC for the purpose of identifying distribution features that can serve as the reference basis for parameterizing the numerical experiments. The one-year hourly EC and OC measurement data from three sites in the PRD (one suburban site in Guangzhou, a general urban site and a roadside site in Hong Kong, with more than 7000 data at each site), are plotted in Figure S1 in the supplemental information (SI) document for the whole year datasets and Figures S2-S4 for the seasonal subsets using the Nansha site as the example. A brief account of the field ECOC analyzers and their field operation is provided in the SI document. A detailed description of the measurement results and data interpretation for the sites will be given in a separate paper. The distributions of measured OC, EC and OC/EC are fitted by both normal and log-normal distribution curves and then examined by the Kolmogorov–Smirnov (K-S) test. The K-S statistic, D, indicates that log-normal fits all three distributions better than the normal distribution (D values are shown in Figure S1-S4). Therefore, log-normal distributions are adopted to define the OC, EC and OC/EC distributions during data generation in our numerical experiments. Statistics of these ambient OC and EC, along with a few other measurements reported in the literature, are summarized in Table 2 and are considered as the reference for data generation to better represent the real situation.

The probability density function (PDF) for the log-normal distribution of variable x is:

$$f(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \times e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}} \tag{3}$$

The two parameters, μ and σ, of the log-normal PDF are related to the average and standard deviation of x through the following equations:

$$\mu = \ln(avg) - 0.5 \times \ln(1 + \frac{std^2}{avg^2}) \tag{4}$$

$$\sigma = \sqrt{\ln(1 + \frac{std^2}{avg^2})} \tag{5}$$

First, realistic average and standard deviation values of EC, $(OC/EC)_{pri}$, and OC (e.g. Figure S1 – S5) are adopted to calculate μ and σ. Then pseudo random number generator use μ and σ to synthesize EC and OC data sets.

The Mersenne twister (MT) (Matsumoto and Nishimura, 1998), a pseudorandom number generator, is used in data generation. MT is provided as a function in Igor Pro. The system clock is utilized as the initial condition for generation of pseudorandom numbers. The data generated by MT has a very long period of $2^{19937} - 1$, permitting large data size and ensuring that pseudorandom numbers are statistically independent between each data generation. The latter feature ensures the independent relationship between EC and non-combustion related SOC data. The case with combustion-related SOC is briefly discussed in section 3. MT also allows assigning a log-normal distribution during pseudorandom number generation to constrain the data. For the verification of the log-normality of MT generated data, a series of K-S tests on the generated data for 5000 runs are conducted. As shown in Figure S6, 94.4% of runs pass the K-S test. Hence the performance of MT can satisfy the log-normal distributed data generation requirement in this study. In a previous study, Chu (2005) used a variant of sine functions to simulate POC and EC, which limited the data size to 120, and the frequency distributions of POC and EC exhibited multiple peaks, a characteristic that is not realistic for ambient measurements. The key information utilized in the EC tracer method is the correlation between EC and POC as well as the irrelevance between EC and SOC. The time series information is not needed in EC tracer method, making pseudorandom number generator a good fit for the evaluation purpose.

The procedure of data generation for the single emission source scenario is illustrated in Figure 2 and implemented by scripts written in Igor Pro. EC is first generated with the following parameters specified: sample size (*n*), average and relative standard deviation (RSD%) of the whole data set (see SI). The EC dataset statistically follows a log-normal distribution, while the sequence of each data point is randomly assigned. POC is then calculated by multiplying EC by $(OC/EC)_{pri}$ (Eq. 1). For simplicity, $(OC/EC)_{pri}$ is set to be a single value, while an analysis incorporating randomly generated log-normally distributed $(OC/EC)_{pri}$ values can be found in the SI material, and a brief summary is given in section 2.2. SOC data is independently generated in a similar way to that for EC. The sum of POC and SOC then yields the synthesized OC. OC and EC data generated in this way are used to calculate SOC by different implementations of the EC tracer method. The bias of SOC estimation can then be evaluated by comparing the calculated SOC with the 'true' SOC values. Data generation for the scenarios with two primary emission sources is similar to the single source scenario and the steps are illustrated in Figure S7.

2.2 Scenario Study

Three scenarios are considered. Scenario 1 (S1) considers one single primary emission source. Scenario 2 (S2) considers two correlated primary emission sources, i.e., two sets of EC, POC, and each source has a single but different $(OC/EC)_{pri}$ value. An example of S2 is combined vehicular emissions from diesel-fuel and gasoline-fuel vehicles. These two sources of vehicular emissions have different $(OC/EC)_{pri}$, but often share a similar temporal variation pattern, making them well correlated. Scenario 3 (S3) considers two independent primary emission sources and simulates an ambient environment influenced by two independent primary emission sources, e.g. local vehicular emissions (lower $(OC/EC)_{pri}$) and regional biomass burning (higher $(OC/EC)_{pri}$).

In the following numerical experiments, three $(OC/EC)_{pri}$ estimation methods are examined and compared, including MRS, $OC/EC_{10\%}$ and OC/EC_{min}. As a single point, OC/EC_{min}, in ambient samples may be subjected to large random uncertainties, thus data with the lowest 1% OC/EC are adopted instead to derive the OC/EC_{min}.

2.2.1 Single primary source scenario

Both $OC/EC_{10\%}$ and OC/EC_{min} methods rely on a subset of ambient OC and EC data to approximate $(OC/EC)_{pri}$. Figure 3 provides a conceptual illustration of the relationships between $(OC/EC)_{pri}$ and the ambient OC/EC data, both are described to exhibit a log-normal distribution. As primary emissions move away from sources and aging processes start in the atmosphere, SOC is added to the particle OC fraction, elevating OC/EC above $(OC/EC)_{pri}$. This in effect broadens the OC/EC distribution curve and shifts the distribution to the right along the OC/EC axis, and the degree of broadening and shift depends on degree of aging process. The conventional EC tracer method using $OC/EC_{10\%}$ and OC/EC_{min} assumes that the left tail of ambient OC/EC distribution is very close to $(OC/EC)_{pri}$. This assumption, however, is fortuitous, rather than the norm. Two parameters, the distance between the means of the $(OC/EC)_{pri}$ and ambient OC/EC distributions and the relative breadth of the two distributions, largely determines the closeness of the approximation of $OC/EC_{10\%}$ and OC/EC_{min} to $(OC/EC)_{pri}$. The distance between the two distributions depends on the fraction of SOC in OC (i.e., f_{SOC}), while the width of the ambient OC/EC distribution is closely associated with RSD of SOC (RSD_{SOC}) and the width of the $(OC/EC)_{pri}$ distribution is reflected in RSD_{POC} and RSD_{EC}. As shown in Figure 3a, only an appropriate combination of distance of the two distribution means and variances could lead to a close approximation of the $(OC/EC)_{pri}$ by $OC/EC_{10\%}$ or OC/EC_{min} (i.e., the left tail of OC/EC distribution). If the ambient aerosol has a significant f_{SOC} shifting the ambient OC/EC distribution such that its left tail is beyond $(OC/EC)_{pri}$ (Figure3b), then the left tail would overestimate $(OC/EC)_{pri}$. Underestimation of $(OC/EC)_{pri}$ could also happen in theory as shown in Figure 3c if the ambient minimum OC/EC (left tail) is less than the mean of the $(OC/EC)_{pri}$ distribution (i.e., under conditions of very small f_{SOC}).

The above analysis reveals f_{SOC}, RSD_{SOC}, RSD_{POC}, and RSD_{EC} are key parameters in influencing the accuracy of SOC estimation. As a result, they are chosen in the subsequent sensitivity tests in probing the SOC estimate bias under conditions of different carbonaceous aerosol compositions.

SOC estimation bias in S1 as a function of RSD_{SOC} and RSD_{EC} are shown in Figures 4a and 4b. The SOC estimate by MRS is not affected by the magnitude of RSD_{EC} and RSD_{SOC}, and is in excellent agreement with the true values (Figure 4). In comparison, SOC by $OC/EC_{10\%}$ and OC/EC_{min} is consistently biased lower and the degree of negative bias becomes larger with decreasing RSD_{SOC} or RSD_{EC}. The $OC/EC_{10\%}$ method always produces larger negative bias than the OC/EC_{min} method. At RSD_{SOC} and RSD_{EC} at 50%, SOC estimate has a -14% bias by $(OC/EC)_{min}$ and a -45% bias by $OC/EC_{10\%}$. These results confirm the hypothesis illustrated in the conceptual diagram (Figure 3) that the validity of using the left tail of OC/EC distribution depends on the distance of its distribution mean from $(OC/EC)_{pri}$ and the distribution breadth. Both $OC/EC_{10\%}$ and the OC/EC_{min} methods underestimate SOC and the degree of underestimation by the $OC/EC_{10\%}$ method is worse.

For the representation of $(OC/EC)_{pri}$ in the simulated data as lognormally distributed data, analysis is also performed to evaluate SOC estimation bias as a function of RSD_{EC}, RSD_{SOC}, and f_{SOC}. Table S2 summarizes the results obtained with adopting most probable ambient conditions (i.e., RSD_{EC}: 50-100%, f_{SOC}: 40-60%). SOC bias by MRS is within 4% when measurement uncertainty is ignored. In comparison, SOC bias by OC/EC_{min} is more sensitive to assumption of log-normally distributed $(OC/EC)_{pri}$ than single value $(OC/EC)_{pri}$, including the dependency on RSD_{EC} and RSD_{SOC} with varied f_{SOC}.

2.2.2 Scenarios assuming two primary sources

In the real atmosphere, multiple combustion sources impacting a site is normal. We next evaluate the performance of the MRS method in scenarios of two primary sources and arbitrarily dictate that the $(OC/EC)_{pri}$ of source 1 is lower than source 2. By varying f_{EC1} (proportion of source 1 EC to total EC) from test to test, the effect of different mixing ratios of the two sources can be examined. Common configurations in S2 and S3 include: EC_{total}=2±0.4 μgC m^{-3}; f_{EC1} varies from 0 to 100%; ratio of the two OC/EC_{pri} values ($\gamma_{\_pri}$) vary in the range of 2~8.

In Scenario 2 (i.e., two correlated primary sources), three factors are examined, including f_{EC1}, $\gamma_{\_pri}$ and f_{SOC}, to probe their effects on SOC estimation. By varying f_{EC1}, the effect of different mixing ratios of two sources can be examined, as f_{EC1} is expected to vary within the same ambient dataset as a result of spatiotemporal dynamics of air masses. MRS reports unbiased SOC, irrespective of different f_{EC1} and f_{SOC} or $\gamma_{\_pri}$ (Figure 5). In comparison, SOC by OC/$EC_{10\%}$ and OC/EC_{min} are underestimated. The degree of underestimation depends on f_{SOC}, e.g., -12% at f_{SOC} = 25% versus -20% at f_{SOC} =40% in the OC/EC_{min} method while the magnitude of underestimation has a very weak dependence on f_{SOC} in the OC/$EC_{10\%}$ method, staying around -40 % as f_{SOC} is doubled from 20% to 40%. The degree of SOC bias by OC/$EC_{10\%}$ and OC/EC_{min} are independent of f_{EC1} and $\gamma_{\_pri}$, as SOC bias is associated with RSD_{EC}, RSD_{SOC} and f_{SOC}. Since two primary sources are well correlated, RSD_{EC} is equivalent between the two sources. As a result, the overall RSD_{EC} is constant when f_{EC1} and $\gamma_{\_pri}$ vary, and the SOC bias is independent of f_{EC1} and $\gamma_{\_pri}$

In summary, in scenarios of two well-correlated primary combustion sources, MRS always produces unbiased SOC estimates while OC/EC_{min} and OC/$EC_{10\%}$ consistently underestimate SOC, with OC/$EC_{10\%}$ producing larger negative bias.

As for Scenario 3 in which two independent primary sources co-exist, SOC estimates by MRS could be biased and the degree and direction of bias depends on f_{EC1}. Figure 6a shows the variation of SOC bias with f_{EC1} when f_{SOC} is fixed at 40%. The variation of SOC bias by MRS with f_{EC1} follows a pseudo-sine curve, exhibiting negative bias when f_{EC1}< 50% (i.e., EC is dominated by source 2, the higher $(OC/EC)_{pri}$ source) and positive bias when f_{EC1}> 50% and the range of bias are confined to -20% to -40% under the condition of f_{SOC}=40%. In comparison, the OC/EC_{min} and OC/$EC_{10\%}$ methods again consistently underestimate SOC by more than -50%, with the bias worsened in the OC/$EC_{10\%}$ method.

The bias variation range becomes narrower with increasing f_{SOC} in the MRS method, as shown by the boxplots for four f_{SOC} conditions (20%, 40%, 60%, and 80%) in Figure 6b. The MRS-derived SOC bias range is reduced from -20–+40% at f_{SOC} = 40% to -10–+20% at f_{SOC} = 60%, further to -6–+10% at f_{SOC} = 80%. In the other two methods, the SOC bias does not improve with increasing f_{SOC}. Dependence of the SOC estimation bias on $\gamma_{\_pri}$ is examined in Figure 6c showing the higher $\gamma_{\_pri}$ induces a higher amplitude of the SOC bias. If OC is dominated by SOC (e.g., f_{SOC} =80%), SOC bias by MRS is within 10%.

A variant of MRS implementation (denoted as MRS') is examined, with the important difference that EC_1 and EC_2, attributed to source 1 and source 2, respectively, are used as inputs instead of total EC. With the knowledge of EC breakdown between the two primary sources, $(OC/EC)_{pri1}$ can be determined by MRS from EC_1 and OC_{total}. Similarly $(OC/EC)_{pri2}$ can be calculated by MRS from EC_2 and OC_{total}. SOC is then calculated with the following equation:

$$SOC = OC_{total} - (OC/EC)_{pri1} \times EC_1 - (OC/EC)_{pri2} \times EC_2 \tag{6}$$

MRS' produces unbiased SOC, irrespective of the different carbonaceous compositions (Figure 6). However, we note that there is a great challenge in meeting the data needs of MRS' as EC_1 and EC_2 are not available.

In scenario 3, the simulation results imply that three factors are associated with the SOC bias by MRS, including: f_{EC1}, $\gamma_{\_pri}$ and f_{SOC}. The first factor controls whether SOC bias by MRS is positive or negative. The latter two affect the degree of SOC bias. For high f_{SOC} conditions, the bias could be acceptable. If EC_1 and EC_2 can be differentiated for calculating individual $(OC/EC)_{pri}$ of each source, unbiased SOC estimation is achievable regardless of what values f_{EC1}, $\gamma_{\_pri}$ and f_{SOC} take.

2.3 Impact of measurement uncertainty

In the preceding numerical analysis, the simulated EC and OC are not assigned any measurement uncertainty; however, in reality, every EC and OC measurement is associated with a certain degree of measurement uncertainty. We next examine the influence of OC and EC measurement uncertainty on SOC estimation accuracy by different EC tracer methods. Two uncertainty types are tested, i.e., constant relative uncertainty (Case A); constant absolute uncertainty (Case B). This section mainly focuses on sensitivity tests assuming different degree of Case A uncertainties. Results assuming Case B uncertainties are discussed in the next section. The uncertainties are assumed to follow a uniform distribution and generated separately by MT. It is also assumed that the uncertainty (ε_{EC} or ε_{OC}) is proportional to the concentration of EC and OC through the multiplier γ_{unc} (i.e., relative measurement uncertainty).

$$-\gamma_{unc}EC \leq \varepsilon_{EC} \leq \gamma_{unc}EC \tag{7}$$

$$-\gamma_{unc}OC \leq \varepsilon_{OC} \leq \gamma_{unc}OC \tag{8}$$

In order to compare the estimated SOC with simulated SOC with ε_{SOC}, the measurement uncertainties of POC and SOC are then back-calculated following the uncertainty propagation formula and assuming the same relative measurement uncertainty for POC and SOC (Harris, 2010)

$$\gamma'_{unc} = \gamma_{unc}\sqrt{\frac{OC^2}{POC^2+SOC^2}} \quad (9)$$

$$-\gamma'_{unc}POC \leq \varepsilon_{POC} \leq \gamma'_{unc}POC \quad (10)$$

$$-\gamma'_{unc}SOC \leq \varepsilon_{SOC} \leq \gamma'_{unc}SOC \quad (11)$$

The simulated EC, POC and SOC with measurement uncertainties (abbreviated as $EC_{simulated}$, $POC_{simulated}$ and $SOC_{simulated}$ respectively) are determined as:

$$EC_{simulated} = EC_{true} + \varepsilon_{EC} \quad (12)$$

$$POC_{simulated} = POC_{true} + \varepsilon_{POC} \quad (13)$$

$$SOC_{simulated} = SOC_{true} + \varepsilon_{SOC} \quad (14)$$

Sensitivity tests of SOC estimation as a function of relative measurement uncertainty (γ_{unc}) and f_{SOC} is performed as shown in Figure 7 by comparing the estimated SOC with $SOC_{simulated}$. Fixed input parameters include: n=8000; EC = 2±1 μgC m^{-3}; $(OC/EC)_{pri}$ = 0.5. Studies by Chu (2005) and Saylor et al. (2006) both suggest ratio of average POC to average EC (ROA, see SI for details) is the best estimator of the expected primary OC/EC ratio because it is mathematically equivalent to the true regression slope when the data contains no intercept. ROA is confirmed as the best representation of $(OC/EC)_{pri}$ for SOC estimation, which shows no bias towards γ_{unc} or f_{SOC} change. MRS overestimates SOC and the positive bias increases with γ_{unc} while decreases with f_{SOC} (Figure 7). The SOC estimates by OC/EC_{min} and $OC/EC_{10\%}$ exhibit larger bias than those by MRS. For example, as shown in Figure 7a, when f_{SOC} =20% and γ_{unc} = 10%, the bias of SOC by MRS, $OC/EC_{10\%}$ and OC/EC_{min} is 8%, -28% and 36%, respectively. With increasing f_{SOC}, the bias of SOC by OC/EC_{min} decreases while the bias of SOC by $OC/EC_{10\%}$ increases when γ_{unc} = 10-20%. MRS always demonstrates the best performance in SOC determination amongst the three $(OC/EC)_{pri}$ estimation methods. When γ_{unc} could be controlled within 20%, the SOC bias by MRS does not exceed 23% when f_{SOC}=20%. If the f_{SOC} ratio falls in the range of 60-80% and γ_{unc} is <20%, the OC/EC_{min} has a similar performance as MRS, but SOC by $OC/EC_{10\%}$ still shows a large bias (~41%) (Figures 7c and 7d).

Sensitivity studies of SOC estimation as a function of γ_{unc} and $(OC/EC)_{pri}$ are performed and the results are shown in Figure S8. In all the three $(OC/EC)_{pri}$ representations, SOC estimates are sensitive to γ_{unc} but insensitive to the magnitude of $(OC/EC)_{pri}$. In the single primary source scenario (S1), it is proved that the performance of MRS regarding SOC estimation is mainly affected by γ_{unc} and to a less degree by f_{SOC}. Other variables such as $(OC/EC)_{pri}$ and EC concentration do not affect the accuracy of SOC estimation.

2.4 Impact of sample size

MRS relies on correlations of input variables and it is expected that MRS performance is sensitive to the sample size of input dataset. This section examines the sensitivity on sample size by the three $(OC/EC)_{pri}$ representations and aims to provide suggestions for an appropriate sample size when applying MRS on ambient OCEC data. Sample sizes ranging from 20 ~ 8000 are tested and for each sample size 500 repeat runs are conducted to obtain statistically significant results. Both Case A (i.e., a constant relative uncertainty of 10%) and Case B (i.e., a constant absolute uncertainty of ±0.2 μgC m^{-3} for both OC and EC) are considered. The measurement uncertainties in case B are generated separately by MT following a uniform distribution within the range of ±0.2 μgC m^{-3}. The measurement uncertainties of POC and SOC are then back-calculated following the uncertainty propagation formula (Harris, 2010) and assuming the ratio of ε_{POC} /ε_{SOC} is the same as POC/SOC ratio (controlled by f_{SOC}).

The mean SOC bias by MRS is very small (<3%) for all sample sizes while the standard deviation of SOC bias decreases with sample size (Figure 8). The standard deviation of SOC bias is ~±30% at the lowest test sample size (n = 20), and decreases to less than ±15% at n = 60 (the sample size of one-year sampling from an every-six-day sampling program) and to less than ±10% at n = 200. Similar patterns are observed between Case A (Figure 8a) and Case B (Figure 8b) for MRS and $OC/EC_{10\%}$. For OC/EC_{min}, a larger bias is observed in Case B than Case A for all sample sizes, as SOC bias by OC/EC_{min} is more sensitive to measurement uncertainty in the range of 0~10% as shown in Figure 7b. The standard deviation of SOC bias by OC/EC_{min} and $OC/EC_{10\%}$ both decreases with sample size as shown in Figure 8. The mean SOC bias of OC/EC_{min} decrease with increased sample size while $OC/EC_{10\%}$ is insensitive to sample size. The sample size dependency of all three $(OC/EC)_{pri}$ representations is not sensitive to f_{SOC} as shown in Figure S16. Other scenarios considering $(OC/EC)_{pri}$ with a distribution and different f_{SOC} are discussed in SI.

2.5 Impact of sampling time resolution

Besides hourly measurements of EC and EC by online aerosol carbon analyzers, the MRS method could also be applied to offline measurements of OC and EC based on filters collected over longer durations (i.e., 24 h), which are more readily available around the world.

To explore the impact of sampling duration (e.g., hourly vs. daily), we here use one-year hourly data at the suburban site of Guangzhou to average them into longer intervals of 2-24 h. The 24 h-averaged samples yield a $(OC/EC)_{pri}$ of 2.53, 12% higher than the $(OC/EC)_{pri}$ derived from hourly data (2.26). This comes as a result of that OC/EC distributions are narrowed when the averaging interval lengthens (Figure 8), leading to elevation of the MRS-derived $(OC/EC)_{pri}$. As many $PM_{2.5}$ speciation networks adopt a sampling schedule of one 24-h sample every six days, we further extract the every-six-day samples to do the MRS calculation. The one-year data yields six subsets of daily samples, corresponding to six possible schedules of sampling days with the every-six-day sampling frequency. The MRS calculation produces the OC/EC_{pri} in the range of 2.37 – 2.75 (5-22% higher than the OC/EC_{pri} from the hourly data). This example illustrates that if 24-h sample ECOC data are used, SOC would be biased slightly lower in comparison with those derived from the hourly data.

3 Caveats of the MRS method in its applications to ambient data

Table 3 summarizes the performance in terms of SOC estimation bias by the different implementations of the EC tracer method, assuming typical variation characteristics for ambient ECOC data. When employing the EC tracer method on ambient samples, it is clear that MRS is preferred since it can provide more accurate SOC estimation.

If the sampling site is dominated by a single primary source (similar to Scenario 1), MRS can perform much better than the traditional OC/EC percentile and minimum approaches. Two issues should be paid attention to when applying MRS: (1) MRS relies on the independence of EC and SOC. This assumption could be invalid if a fraction of SOC is formed from semi-volatile POC (here referred as SOC_{svP}) (Robinson et al., 2007). Since POC is well correlated with EC, this SOC_{svP} would be attributed to POC by MRS, causing SOC underestimation. The interference of SOC_{svP} will be discussed in a separate paper. (2) $OC_{non\text{-}comb}$ will be attributed to SOC if only EC is used as a tracer. If $OC_{non\text{-}comb}$ is small compared to SOC, such approximation is acceptable. Otherwise quantification of its contribution is needed. If a stable tracer for $OC_{non\text{-}comb}$ is available, determination of $OC_{non\text{-}comb}$ contribution by MRS is possible, since this scenario is mathematically equivalent to S3 (e.g., relabel EC2 to tracer of $OC_{non\text{-}comb}$ and POC to $OC_{non\text{-}comb}$).

If the sampling site is influenced by two correlated primary sources with distinct $(OC/EC)_{pri}$ (Scenario 2, e.g. urban areas that have vehicular emission from both gasoline and diesel), MRS is still much more reliable than the traditional OC/EC percentile and minimum approaches. If the sampling site is influenced by two independent primary sources with distinct $(OC/EC)_{pri}$ (Scenario 3, e.g. vehicular emission and biomass burning), SOC estimation by MRS is better than the other two conventional methods. But it should be noted that possible bias may exist and the magnitude of bias depends on the relative abundance between the two sources. If tracers are available to demarcate the EC contributions by the different primary sources, unbiased SOC estimation is possible by employing these tracers in MRS.

4 Conclusions

In this study, the accuracy of SOC estimation by EC tracer method is evaluated by comparing three $(OC/EC)_{pri}$ determination approaches using numerically simulated data. The MRS method has a clear quantitative criterion for the $(OC/EC)_{pri}$ calculation, while the other two commonly used methods, namely minimum OC/EC (OC/EC_{min}) and OC/EC percentile (e.g. $OC/EC_{10\%}$), are empirical in nature. Three scenarios are considered in the numerical simulations to evaluate the SOC estimation bias by the different EC tracer methods assuming typical variation characteristics for ambient ECOC data. In the scenarios of a single primary source and two well-correlated primary combustion sources, SOC estimates by MRS are unbiased while OC/EC_{min} and $OC/EC_{10\%}$ consistently underestimate SOC when measurement uncertainty is neglected. When measurement uncertainty is considered, all three approaches produce biased SOC estimates, with MRS producing the smallest bias. The bias by MRS is less than 25% when measurement uncertainty is within 20% and f_{SOC} is not lower than 20%. In the scenario of two independent primary sources, SOC by MRS exhibit bias but still perform better than OC/EC_{min} and $OC/EC_{10\%}$. If EC from each independent source can be differentiated to allow calculation of individual $(OC/EC)_{pri}$ for each source, unbiased SOC estimation is achievable. Sensitivity tests of OC and EC measurement uncertainty on SOC estimation demonstrate the superior accuracy of MRS over the other two approaches.

Sensitivity tests show that MRS produces a mean SOC values with a very small bias for all sample sizes while the precision worsens as the sample size decreases. For a dataset with a sample size of 60, SOC bias by MRS is 2±15%. When the sample is 200, the results by MRS are improved to 2±8%. It is clear that when employing the EC tracer method to estimate SOC, MRS is preferred over the two conventional methods ($OC/EC_{10\%}$ and OC/EC_{min}) since it can provide more accurate SOC estimation. We also evaluated the impact of longer sampling duration on derived $(OC/EC)_{pri}$ and found that if 24-h sample ECOC data are used, SOC would be biased slightly lower in comparison with those derived from the hourly data.

Supporting Information

The Supplement related to this article is available online.

Acknowledgements

This work is supported by the National Science Foundation of China (21177031), and the Fok Ying Tung Graduate School (NRC06/07.SC01). The authors thank Hong Kong Environmental Protection Department for making available the ECOC data at Tsuen Wan and Prof. Dui Wu of Institute of Tropical and Marine Meteorology, China Meteorological Administration for providing logistic support of OC EC measurements in Nancun. The authors are also grateful to Dr. Stephen M Griffith for the helpful comments.

References

Batmunkh, T., Kim, Y. J., Lee, K. Y., Cayetano, M. G., Jung, J. S., Kim, S. Y., Kim, K. C., Lee, S. J., Kim, J. S., Chang, L. S., and An, J. Y.: Time-Resolved Measurements of PM2.5 Carbonaceous Aerosols at Gosan, Korea, J Air Waste Manage, 61, 1174-1182, 2011.

Cao, J. J., Lee, S. C., Ho, K. F., Zou, S. C., Fung, K., Li, Y., Watson, J. G., and Chow, J. C.: Spatial and seasonal variations of atmospheric organic carbon and elemental carbon in Pearl River Delta Region, China, Atmos Environ, 38, 4447-4456, DOI 10.1016/j.atmosenv.2004.05.016, 2004.

Castro, L. M., Pio, C. A., Harrison, R. M., and Smith, D. J. T.: Carbonaceous aerosol in urban and rural European atmospheres: estimation of secondary organic carbon concentrations, Atmos Environ, 33, 2771-2781, 1999.

Chu, S. H.: Stable estimate of primary OC/EC ratios in the EC tracer method, Atmos Environ, 39, 1383-1392, 2005.

Gray, H. A., Cass, G. R., Huntzicker, J. J., Heyerdahl, E. K., and Rau, J. A.: Characteristics of atmospheric organic and elemental carbon particle concentrations in Los Angeles, Environ Sci Technol, 20, 580-589, 10.1021/es00148a006, 1986.

Harris, D. C.: Quantitative chemical analysis, 8th ed., W.H. Freeman and Co., New York, 2010.

Hu, D., Bian, Q. J., Lau, A. K. H., and Yu, J. Z.: Source apportioning of primary and secondary organic carbon in summer PM2.5 in Hong Kong using positive matrix factorization of secondary and primary organic tracer data, J Geophys Res-Atmos, 115, D16204, Doi 10.1029/2009jd012498, 2010.

Hu, W. W., Hu, M., Deng, Z. Q., Xiao, R., Kondo, Y., Takegawa, N., Zhao, Y. J., Guo, S., and Zhang, Y. H.: The characteristics and origins of carbonaceous aerosol at a rural site of PRD in summer of 2006, Atmos Chem Phys, 12, 1811-1822, 2012.

Huang, X. H., Bian, Q., Louie, P., and Yu, J.: Contributions of vehicular carbonaceous aerosols to PM 2.5 in a roadside environment in Hong Kong, Atmos Chem Phys, 14, 9279-9293, 2014.

Lim, H. J., and Turpin, B. J.: Origins of primary and secondary organic aerosol in Atlanta: Results' of time-resolved measurements during the Atlanta supersite experiment, Environ Sci Technol, 36, 4489-4496, 2002.

Lin, P., Hu, M., Deng, Z., Slanina, J., Han, S., Kondo, Y., Takegawa, N., Miyazaki, Y., Zhao, Y., and Sugimoto, N.: Seasonal and diurnal variations of organic carbon in PM2.5 in Beijing and the estimation of secondary organic carbon, J Geophys Res-Atmos, 114, D00G11, 10.1029/2008JD010902, 2009.

Malm, W. C., Schichtel, B. A., Pitchford, M. L., Ashbaugh, L. L., and Eldred, R. A.: Spatial and monthly trends in speciated fine particle concentration in the United States, Journal of Geophysical Research: Atmospheres, 109, D03306, 10.1029/2003jd003739, 2004.

Matsumoto, M., and Nishimura, T.: Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator, ACM Trans. Model. Comput. Simul., 8, 3-30, 10.1145/272991.272995, 1998.

Millet, D. B., Donahue, N. M., Pandis, S. N., Polidori, A., Stanier, C. O., Turpin, B. J., and Goldstein, A. H.: Atmospheric volatile organic compound measurements during the Pittsburgh Air Quality Study: Results, interpretation, and quantification of primary and secondary contributions, Journal of Geophysical Research: Atmospheres, 110, D07S07, 10.1029/2004jd004601, 2005.

Pio, C., Cerqueira, M., Harrison, R. M., Nunes, T., Mirante, F., Alves, C., Oliveira, C., de la Campa, A. S., Artinano, B., and Matos, M.: OC/EC ratio observations in Europe: Re-thinking the approach for apportionment between primary and secondary organic carbon, Atmos Environ, 45, 6121-6132, DOI 10.1016/j.atmosenv.2011.08.045, 2011.

Polidori, A., Turpin, B. J., Lim, H. J., Cabada, J. C., Subramanian, R., Pandis, S. N., and Robinson, A. L.: Local and regional secondary organic aerosol: Insights from a year of semi-continuous carbon measurements at Pittsburgh, Aerosol Sci Tech, 40, 861-872, 2006.

Robinson, A. L., Donahue, N. M., Shrivastava, M. K., Weitkamp, E. A., Sage, A. M., Grieshop, A. P., Lane, T. E., Pierce, J. R., and Pandis, S. N.: Rethinking Organic Aerosols: Semivolatile Emissions and Photochemical Aging, Science, 315, 1259-1262, 10.1126/science.1133061, 2007.

Saylor, R. D., Edgerton, E. S., and Hartsell, B. E.: Linear regression techniques for use in the EC tracer method of secondary organic aerosol estimation, Atmos Environ, 40, 7546-7556, DOI 10.1016/j.atmosenv.2006.07.018, 2006.

Turpin, B. J., and Huntzicker, J. J.: Secondary Formation of Organic Aerosol in the Los-Angeles Basin - a Descriptive Analysis of Organic and Elemental Carbon Concentrations, Atmos Environ a-Gen, 25, 207-215, 1991.

Turpin, B. J., and Huntzicker, J. J.: Identification of Secondary Organic Aerosol Episodes and Quantitation of Primary and Secondary Organic Aerosol Concentrations during Scaqs, Atmos Environ, 29, 3527-3544, 1995.

Wang, Z., Wang, T., Guo, J., Gao, R., Xue, L. K., Zhang, J. M., Zhou, Y., Zhou, X. H., Zhang, Q. Z., and Wang, W. X.: Formation of secondary organic carbon and cloud impact on carbonaceous aerosols at Mount Tai, North China, Atmos Environ, 46, 516-527, 2012.

Yu, S. C., Dennis, R. L., Bhave, P. V., and Eder, B. K.: Primary and secondary organic aerosols over the United States: estimates on the basis of observed organic carbon (OC) and elemental carbon (EC), and air quality modeled primary OC/EC ratios, Atmos Environ, 38, 5257-5268, DOI 10.1016/j.atmosenv.2004.02.064, 2004.

Table 1. Acronyms and Abbreviations

Abbreviation	**Definition**
EC	elemental carbon
EC_1,EC_2	EC from source 1 and source 2 in the two sources scenario
f_{EC1}	fraction of EC from source 1 to the total EC
f_{SOC}	ratio of SOC to OC
MRS	minimum R squared method
MRS'	a variant of MRS that use EC from individual sources as input
MT	Mersenne twister pseudorandom number generator
n	sample size in MT data generation
OC	organic carbon
OC/EC	OC to EC ratio
$(OC/EC)_{pri}$	primary OC/EC
$OC/EC_{10\%}$	OC/EC at 10% percentile
OC/EC_{min}	minimum OC/EC
$OC_{non\text{-}comb}$	OC from non-combustion sources
PDF	probability density function of a distribution
POC	primary organic carbon
ROA	ratio of averages
RSD	relative standard deviation
RSD_{EC}	RSD of EC
RSD_{POC}	RSD of POC
RSD_{SOC}	RSD of SOC
SOC	secondary organic carbon
SOC_{svP}	SOC formed from semi-volatile POC
Υ_pri	ratio of the $(OC/EC)_{pri}$ of source 2 to source 1
ε_{EC} , ε_{OC}	measurement uncertainty of EC and OC
Υ_{unc}	relative measurement uncertainty
γ_RSD	the ratio between the RSD values of $(OC/EC)_{pri}$ and EC

Table 2. Summary of statistics of OC and EC in ambient samples

Location	Site Type	Sampling Period	Time resolution	RSD_{EC} (%)	RSD_{SOC} (%)	SOC estimation method	f_{SOC} mass fraction (%)			Ref
							Avg	Min	Max	
Hong Kong, PRD	Suburban	July 2006, local days	24 hr			PMF	25%	6%	79%	(Hu et al., 2010)
		July 2006, regional days					65%	46%	89%	
Hong Kong, PRD	Urban	May 2011 – Apr. 2012	1 hr	51%		EC tracer PMF				(Huang et al., 2014)
Guangzhou, PRD	Rural	July 2006	1 hr	154%	115%	EC tracer	47%		80%	(Hu et al., 2012)
Guangzhou, PRD	Suburban	Feb 2012 – Jan 2013	1 hr	86%	84%	EC tracer	41%	0%	86%	This study
Beijing	Urban	Winter 2005	1 hr			EC tracer	19%			(Lin et al., 2009)
		Spring 2006					27%			
		Summer 2006					45%			
		Fall 2006					23%			
Pittsburgh	Suburban	Jul. 2001 – Aug. 2002	2-4 hr			EC tracer	38%			(Polidori et al., 2006)
Mt. Tai, China	Rural	Mar. – Apr. 2007	1 hr	89%		EC tracer	60%			(Wang et al., 2012)
		Jun. – Jul. 2007		69%			73%			
Jeju Island, Korea	Rural	May – Jun. 2009	1 hr	53%	117%	EC tracer	31%			(Batmunkh et al., 2011)
		Aug – Sep 2009		57%	102%		18%			

Table 3. Summary of numerical study results under different scenarios [a].

	Tested parameter	SOC bias			
		MRS[b]	MRS'[c]	OC/EC_{min}	$OC/EC_{10\%}$
Scenario 1 Single source	RSD_{EC}	±4%		-13% ~ -7%	-43% ~ -36%
	RSD_{SOC}	±4%		-11% ~ -4%	-42% ~ -22%
	γ_{unc}	+10%		-12% ~ 20%	-43% ~ -32%
Scenario 2 Two correlated sources	f_{EC1}	±4%		-20%	-40%
	γ_pri	±4%		-20%	-40%
	f_{SOC}	±4%		-20%	-40%
Scenario 3 Two independent sources	f_{EC1}	-20%~40%	±10%	-50%	-60%
	γ_pri	-20%~40%	±10%	-50%	-60%
	f_{SOC}	-20%~40%	±10%	-50%	-60%

[a] Results shown here are obtained assuming the following ambient conditions: RSD_{EC} 50-100%; f_{SOC} 40-60%; γ_{unc} 20%;
[b] "+" represents SOC overestimation and "-" represents underestimation;
[c] MRS': In S3, EC1 and EC2 are used for SOC calculation.

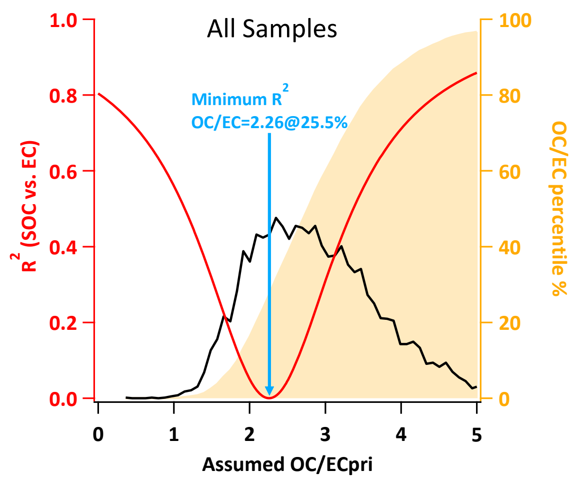

Figure 1. Illustration of the minimum R square method (MRS) to determine OC/EC_{pri} using one year of hourly OC and EC measurements at a suburban site in the Pearl River Delta, China. The red curve shows the correlation coefficient (R^2) between SOC and EC as a function of assumed OC/EC_{pri}. The black curve is the frequency distribution of the OC/EC ratio for the entire OC and EC data set. The shaded area in tan represents the cumulative frequency curve of OC/EC ratio.

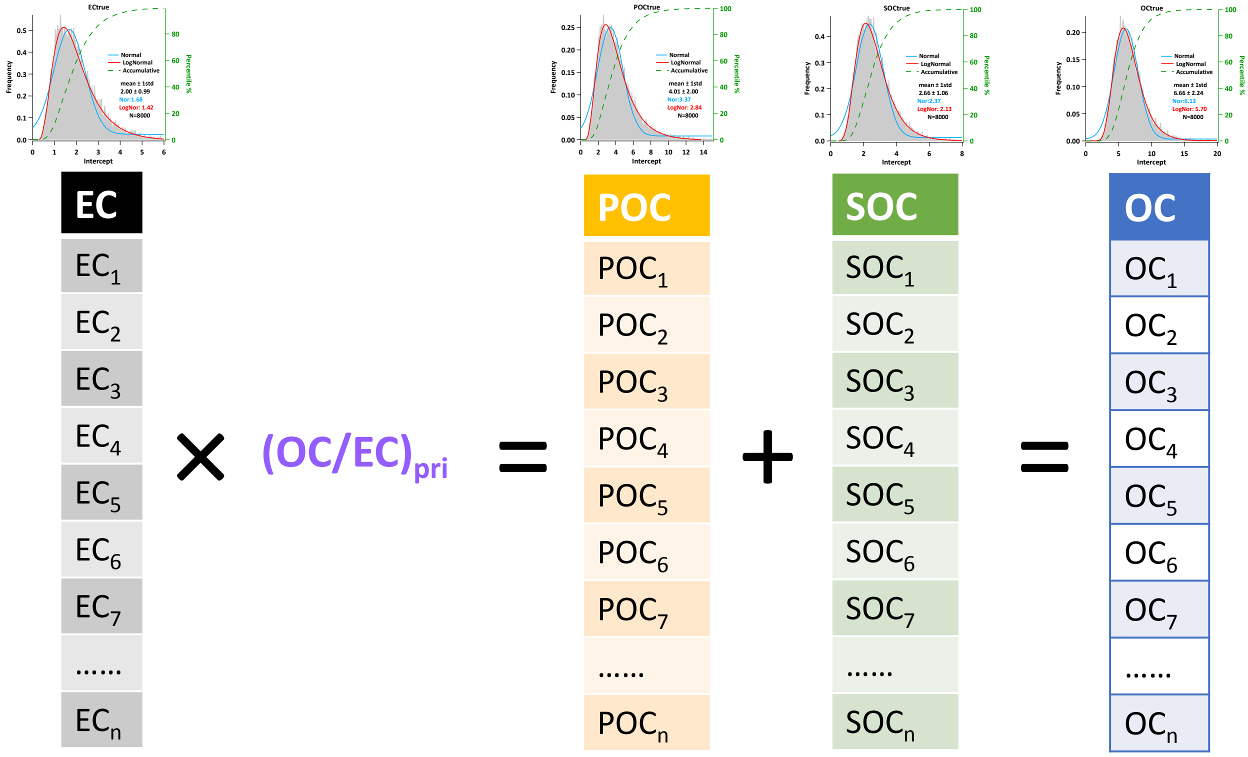

Figure 2. Schematic diagram of pseudorandom number generation for the single emission source scenario that assumes $(OC/EC)_{pri}$ is a single value. The data series (EC and SOC), generated by Mersenne twister (MT) pseudorandom number generator, statistically follow a log-normal distribution, but the sequence of each data point is randomly assigned.

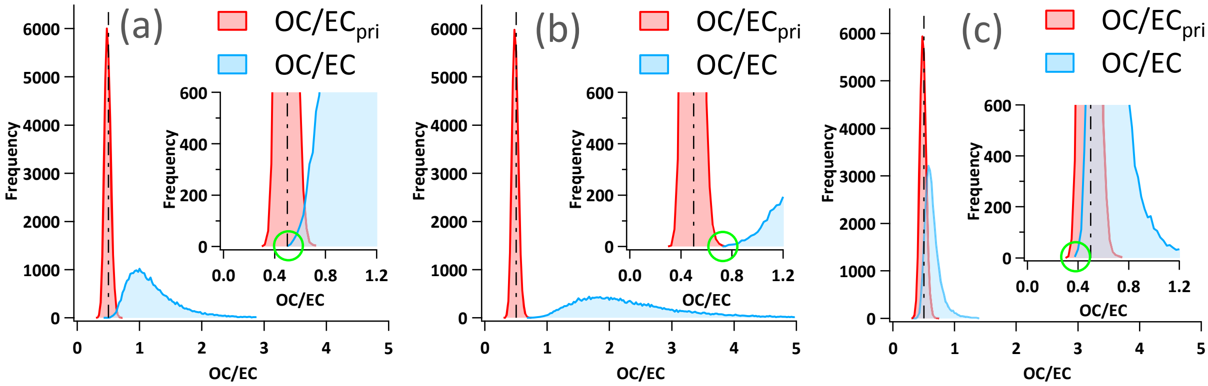

Figure 3. Conceptual diagram illustrating three scenarios of the relationship between $(OC/EC)_{pri}$ and ambient OC/EC measurements. Both are assumed to be log-normally distributed. (a) Ambient minimum (left tail) is equal to the peak of $(OC/EC)_{pri}$. (b) Ambient minimum OC/EC (left tail) is larger than the mean of $(OC/EC)_{pri}$. (c) Ambient minimum OC/EC (left tail) is less than the peak of $(OC/EC)_{pri}$.

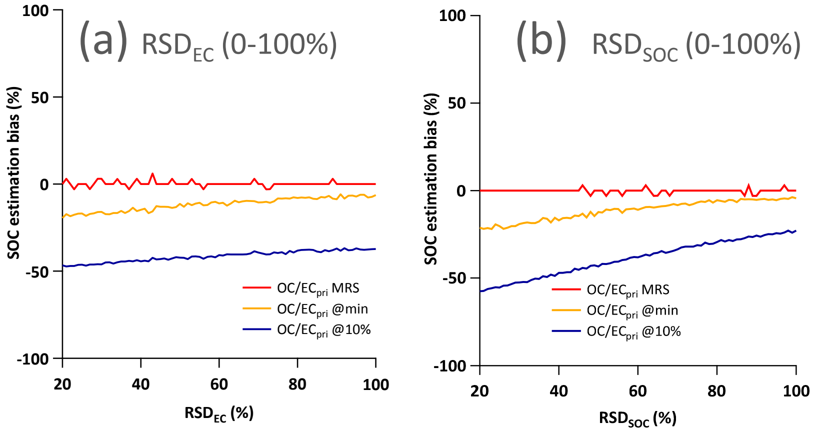

Figure 4. Bias of SOC determination as a function of: (a) RSD_{EC}; (b) RSD_{SOC}. Different representation of $(OC/EC)_{pri}$ include: MRS, OC/EC_{min} and $OC/EC_{10\%}$.. Fixed input parameters: n = 8000, EC = 2±1 μgC m^{-3}, $(OC/EC)_{pri}$ = 0.5, POC = 1 ±0.5 μgC m^{-3}, f_{SOC} =40%, and SOC = 0.67±0.34 μgC m^{-3}.

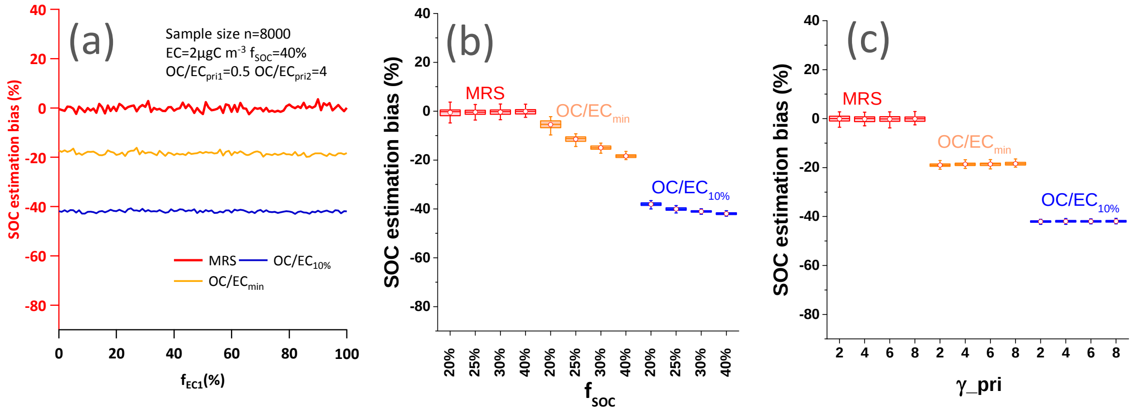

Figure 5. SOC bias in Scenario 2 (two correlated primary emission sources of different $(OC/EC)_{pri}$) as estimated by four different EC tracer methods denoted in red, blue and yellow. (a) SOC bias as a function of f_{EC1}. Results shown here are calculated using f_{SOC} = 40% as an example. (b) Range of SOC bias shown in boxplots for four f_{SOC} conditions (20%, 25%, 30% and 40%). (c) Range of SOC bias shown in boxplots for four γ_pri conditions (2, 4, 6 and 8). The symbols in the boxplots are white circles for average, the line inside the box for median, the box boundaries representing the 75th and the 25th percentile, and the whiskers representing the 95th and 5th percentile.

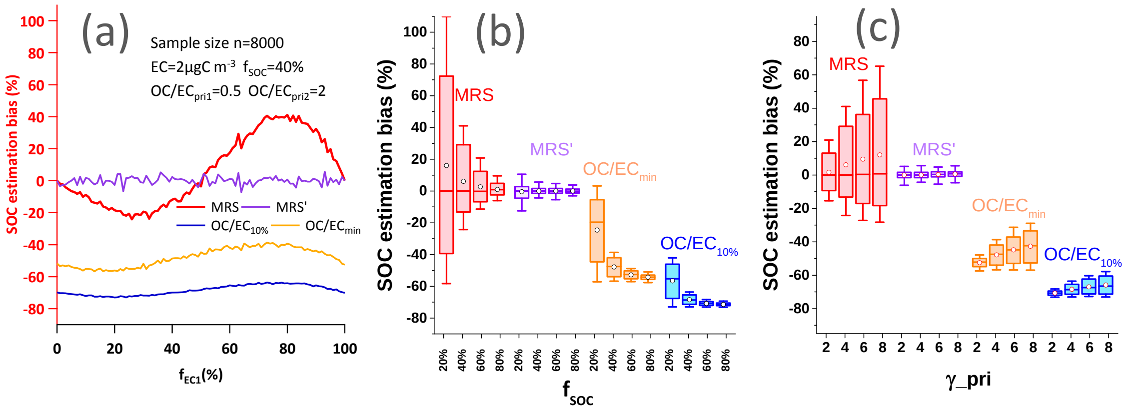

Figure 6. SOC bias in Scenario 3 (two independent primary emission sources of different $(OC/EC)_{pri}$) as estimated by four different EC tracer methods denoted in red, purple, blue and yellow. MRS' differs from MRS in that EC_1 and EC_2 instead of total EC are used as inputs. (a) SOC bias as a function of f_{EC1}. Results shown here are calculated using f_{SOC} = 40% as an example. (b) Range of SOC bias shown in boxplots for four f_{SOC} conditions (20%, 40%, 60% and 80%). (c) Range of SOC bias shown in boxplots for four γ_pri conditions (2, 4, 6 and 8). The symbols in the boxplots are white circles as average, the line inside the box as median, upper and lower boundaries of the box representing the 75th and the 25th percentile, and the whiskers above and below each box representing the 95th and 5th percentile.

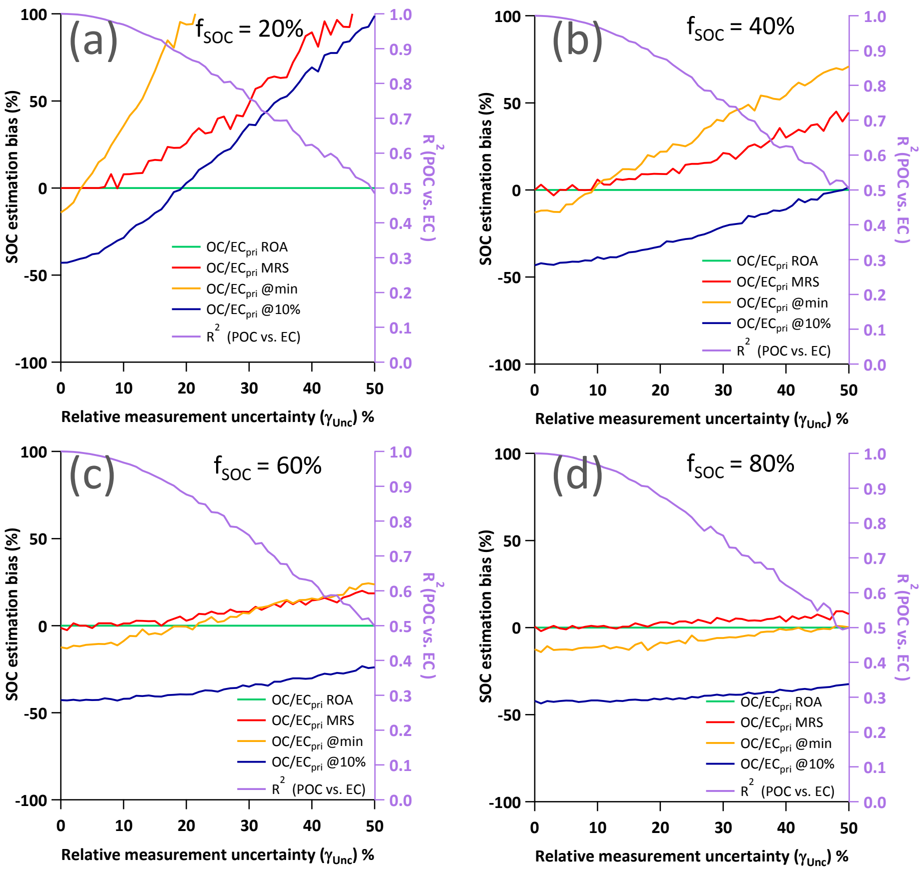

Figure 7. Bias of SOC determination as a function of relative measurement uncertainty (γ_{unc}) and SOC/OC ratio (f_{SOC}) by different approaches of estimating $(OC/EC)_{pri}$, including ratio of averages (ROA), minimum R square (MRS), $OC/EC_{10\%}$, and OC/EC_{min}. Fixed input parameters: *n*=8000, EC = 2±1 μgm^{-3}, $(OC/EC)_{pri}$ = 0.5. Variable input parameters: (a) f_{SOC} =20%, SOC = 0.25±0.13 μgC m^{-3}, (b) f_{SOC} =40%, SOC = 0.67±0.33 μgC m^{-3}, (c) f_{SOC} =60%, SOC = 1.5±0. 75 μgC m^{-3}, and (d) f_{SOC} =80%, SOC = 4±2 μgC m^{-3}

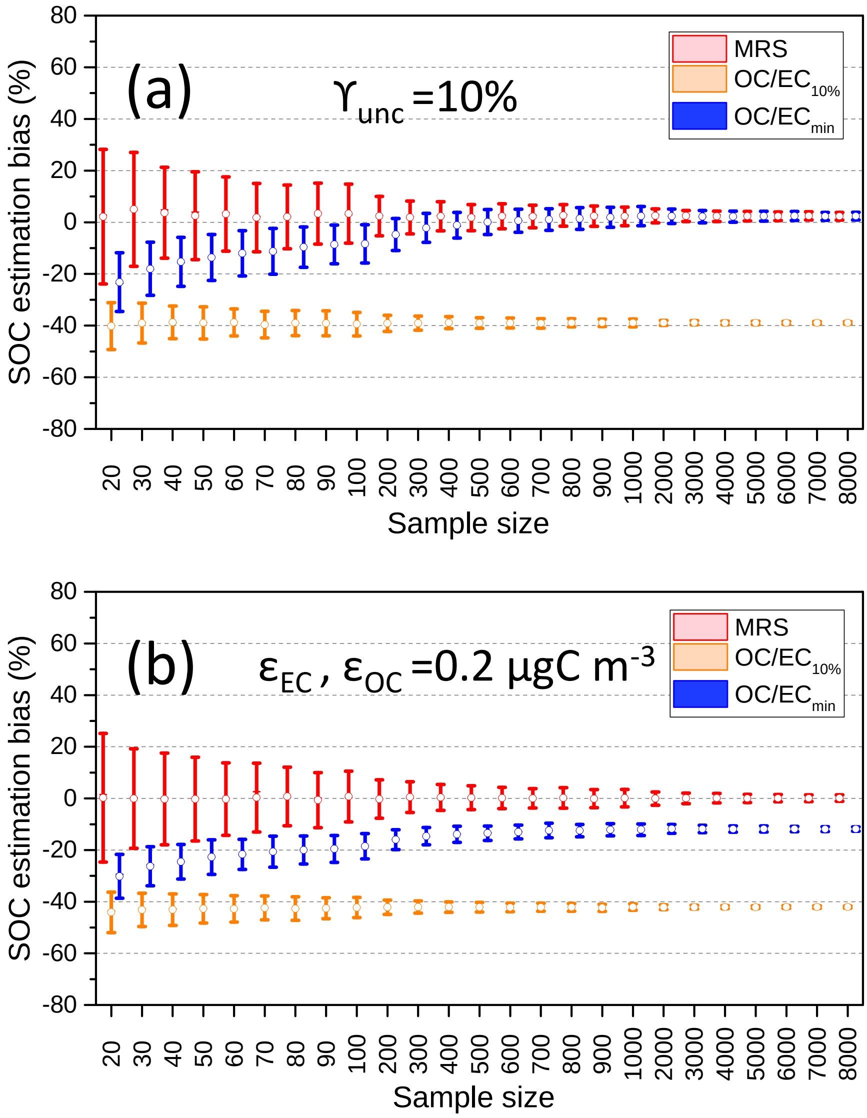

Figure 8. SOC estimation bias as a function of sample size by different approaches of estimating $(OC/EC)_{pri}$, including minimum R square (MRS), $OC/EC_{10\%}$, and OC/EC_{min}, (a) assuming a fixed relative measurement uncertainty of 10% for OC and EC; (b) assuming a fixed absolute measurement uncertainty for OC and EC (0.2 μg m^{-3}). For each sample size, 500 repeat runs were conducted. The circles represent mean of 500 repeat runs, the whiskers represent one standard deviation. Parameters used for testing: Repeat runs = 500, n = 20~8000, EC = 8±4 μgC m^{-3}, $(OC/EC)_{pri}$ = 0.5, POC = 4±2 μgC m^{-3}, f_{SOC} =40%, and SOC = 2.67±1.33 μgC m^{-3}.

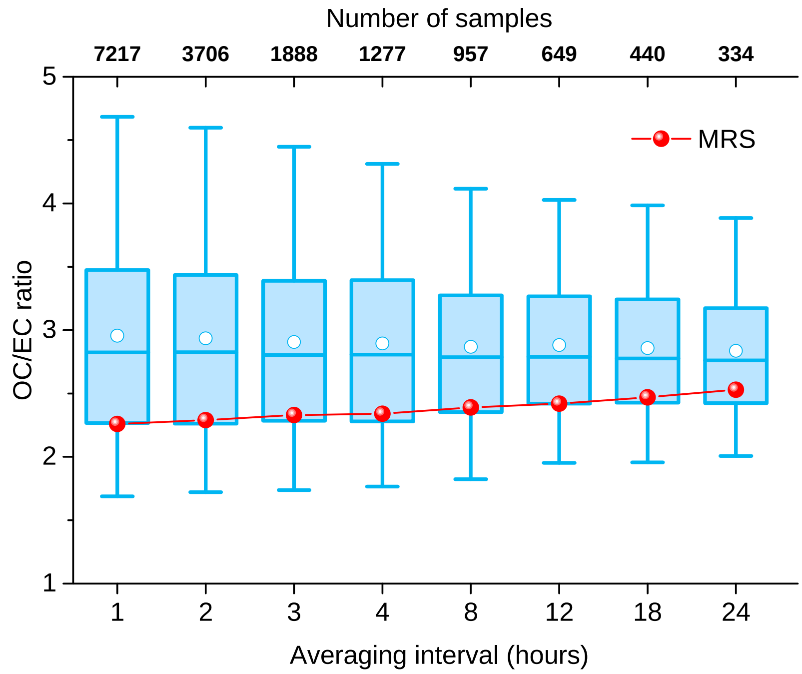

Figure 9. OC/EC distributions assuming different average intervals from 2 to 24 h and the corresponding MRS-derived OC/EC_{pri}. The bottom x-axis represents averaging interval (e.g. 1 h is the original data time resolution, 2 h referring average 1-h data into 2-h interval data, etc). The top x-axis represents the number of data point corresponding to the respective data averaging interval. Distributions of OC/EC ratio at various averaging intervals are shown as box plots (Empty circles: average, the line inside the box: median, the box boundaries: 75^{th} and the 25^{th} percentile, and the whiskers: 95^{th} and 5^{th} percentile). The red dots represent calculated $(OC/EC)_{pri}$ by MRS.