Point-by-point response to review comments on manuscript amt-2017-300 "Evaluation of linear regression techniques for atmospheric applications: The importance of appropriate weighting"

By Cheng Wu et and Jian Zhen Yu

We thank the two anonymous reviewers for their constructive comments to improve the manuscript. 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

R1-Q1.

The paper is an extension of the work by Saylor et al. (2006) and shows that ordinary least squares (OLS) techniques are not the best techniques in comparing two variables which both have errors in measurements.

The paper is well written and the science is good.

However, one can discuss the 'new science' of the paper. What is discussed in the paper, that OLS is a flawed method for comparing variables with errors, should be known to many researchers. However, reviewing the literature, one can see that it is not as widely known as it should be. Indeed, the OLS is often still abused in literature. Therefore, if this paper manages to increase the knowledge in using better regression methods for these cases, it will have served it purpose. As a result, despite the lack of a lot of 'new science', I would still accept the paper, albeit when another case that is lacking now is discussed. Discussion of this case would improve the usefulness of this paper strongly in my opinion: OLS is still widely used when comparing for instance model and measurement data. It would be interesting to add such a case, where the a priori error in one of the variables is unknown. What regression techniques would then be ideal? This can happen too with measurement techniques, if for instance, the technical errors of a measurement described cannot be trusted. And what is the best technique if the errors on both the independent and the dependent variable are unknown? How to proceed in that case?

Adding this discussion would, in my opinion, improve the manuscript.

Author's Response: The reviewer raised a very good point and we fully agree that including corresponding tests would improve the usefulness of the manuscript. To address this question, we added a new section with two tests (Figure R-1) in the manuscript. The corresponding discussion are shown below.

4.4 Caveats of regressions with unknown X and Y uncertainties

When applying linear regression on real world data, it happens that a priori error in one of the variables is unknown, or the measurement error described cannot be trusted. In other words, that would be certain degree of discrepancy between the measurement error used for linear regression and measurement error embed in the data. It is common that measurement error cannot be determined due to the lack of duplicated or collocated measurements and an arbitrarily assumed uncertainty is used. For example, Flanagan et al. (2006) found that the whole-system uncertainty retrieved by data from collocated sampler is different from the arbitrarily assumed 5% uncertainty, which is previously used by the Speciation Trends Network (STN). In addition, the degree of discrepancy between the actual uncertainty by collocated samples and arbitrarily assumed uncertainty also varied by different chemical species.

To investigate the impact of such cases on different regression approaches, two tests are conducted. In Test A, the actual measurement error for X is fixed at 30% while γ_{Unc} for Y varied from 1% to 50%. The assumed measurement error for regression is 10% for both X and Y. Results of Test A are shown in Figure 6 a&b. For OLS, the slopes are underestimated (-14 ~ -12%) and intercepts are overestimated (90 ~ 103%). The biases in OLS slope and intercept are independent of variations in γ_{Unc_Y}. ODR and DR ($\lambda = 1$) yield similar results with overestimated slopes (0 ~ 44%) and underestimated intercepts (-330 ~ 0%). The degree of bias in slopes and intercepts depends on γ_{Unc_Y}. WODR, DR ($\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$) and YR performed much better than other regression approaches in Test A, with a smaller bias in both slopes (-8 ~ 12%) and intercepts -98 ~ 55%).
The results of Test B are shown in Figure 6 c&d. which has a fixed γ_{Unc_Y} of 30% and γ_{Unc_X} varied between 1 ~ 50%. The assumed measurement error for regression is 10% for both X and Y. OLS underestimates slopes (-29 ~- 0.2%) and overestimates intercepts (2 ~ 209%) in Test B. In contrast to Test A which slope and intercept biases are independent of variations in γ_{Unc_Y}, the OLS slope and intercept biases in Test B exhibit dependency on γ_{Unc_X}. The reason behind is because OLS only considers errors in Y, while X is assumed to be error free. ODR and DR ($\lambda = 1$) yield similar results with overestimated slopes (11 ~ 18%) and underestimated intercepts (-144 ~ -87%). The degree of biases in slopes and intercepts is relatively independent to the γ_{Unc_X}. WODR, DR ($\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$) and YR performed much better than other regression approaches in Test B, with a smaller bias in both slopes (-14 ~ 8%) and intercepts (-59 ~ 106%).
The results from these two tests suggest that, in case of one of the measurement error described cannot be trusted or a priori error in one of the variables is unknown, WODR, DR ($\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$) and YR should be used instead of ODR, DR ($\lambda = 1$) and OLS. This conclusion also agrees with section 4.1 and 4.2. The results also suggest that, in general, the magnitude of bias in slope estimation by these regression approaches are smaller than those for intercept. In other words, slope is a more reliable quantity compare to intercept when extracting quantitative information from linear regressions.

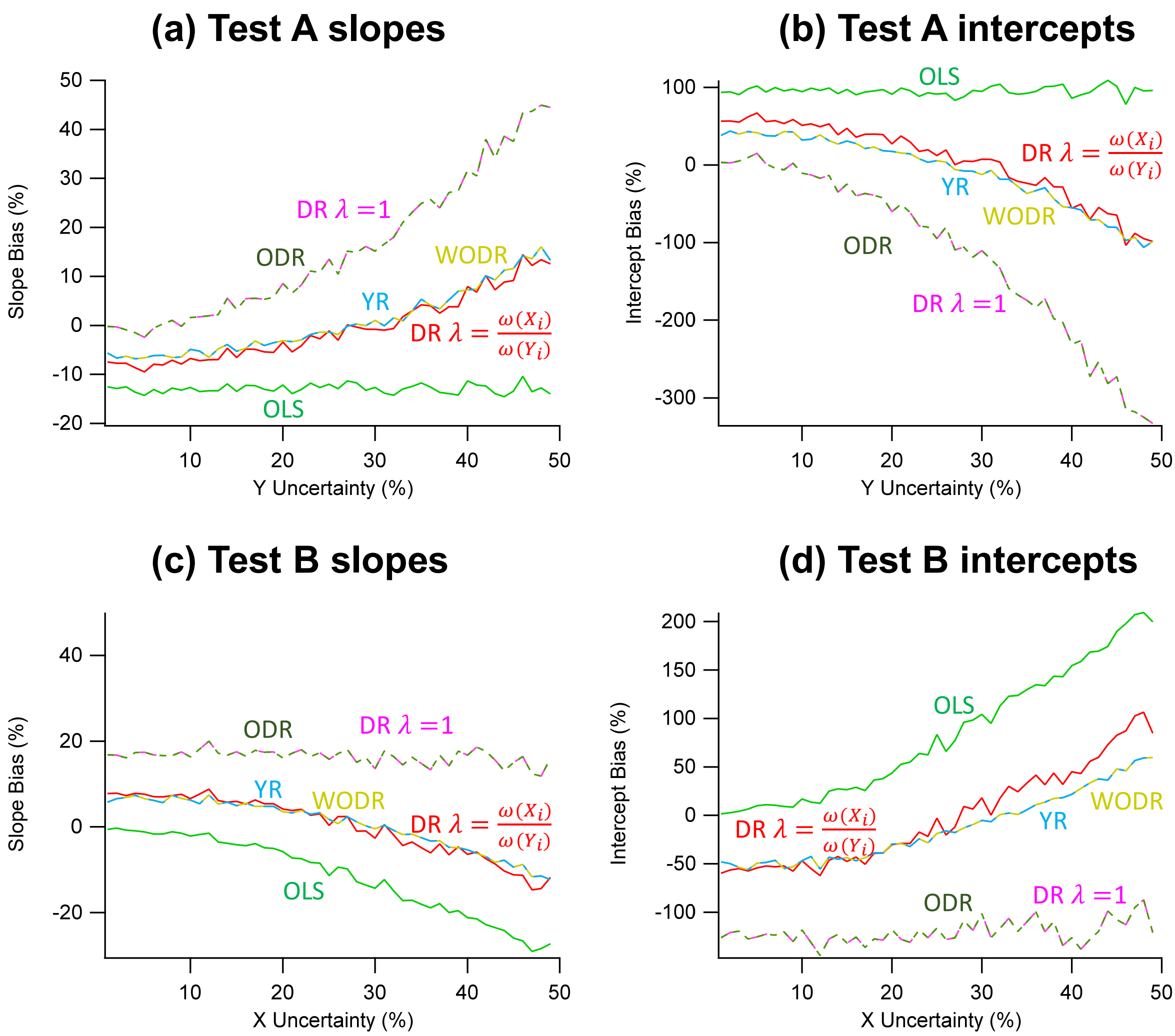


Figure R-1. Slope and intercept biases due to the inconsistency between measurement error of data and measurement error used in regression. In Test A data generation, γ_{Unc_X} is fixed at 30% and γ_{Unc_Y} varied between 1 ~ 50%. In Test B, γ_{Unc_X} varied between 1 ~ 50% and γ_{Unc_Y} is fixed at 30%. The assumed measurement error for regression is 10% for both X and Y. (a) Slopes biases as a function of γ_{Unc_Y} in Test A. (b) Intercepts biases as a function of γ_{Unc_Y} in Test A. (c) Slopes biases as a function of γ_{Unc_X} in Test B. (d) Intercepts biases as a function of γ_{Unc_X} in Test B.

Following contents are added to the abstract to cover the findings in section 4.4.

> If discrepancy exist between measurement error of data and measurement uncertainty used for regression, DR, WODR and YR can provide the least biases in slope and intercept among all tested regression techniques. For these reasons, DR, WODR and YR are recommended for atmospheric studies when both X and Y data have measurement errors.

The first paragraph of conclusion is updated to reflect the finding in section 4.4.

> This study aims to provide a benchmark of commonly used linear regression algorithms using a new data generation scheme (MT). Six regression approaches are tested, including OLS, DR ($\lambda = 1$), DR ($\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$), ODR, WODR and YR.

The results show that OLS fails to estimate the correct slope and intercept when both X and Y have measurement errors. This result is consistent with previous studies. For ambient data with R^2 less than 0.9, error-in-variables regression is needed to minimize the biases in slope and intercept. If measurement uncertainties in X and Y are determined during the measurement, measurement uncertainties should be used for regression. With appropriate weighting, DR, WODR and YR can provide the best results among all tested regression techniques. Sensitivity tests also reveal the importance of the weighting parameter λ in DR. An improper λ could lead to biased slope and intercept. Since the λ estimation depends on the form of the measurement errors, it is important to determine the measurement errors during the experimentation stage rather than making assumptions. If measurement errors are not available from the measurement and assumptions are made on measurement errors, DR, WODR and YR are still the best option that can provide the least bias in slope and intercept among all tested regression techniques. For these reasons, DR, WODR and YR are recommended for atmospheric studies when both X and Y data have measurement errors.

Technical comments

R1-Q2. Last sentence of §3.1.2: meaning of SI?

Author's Response: Supplemental information. The sentence had been revised to "A brief introduction is given in the Supplemental Information."

Evaluation of linear regression techniques for atmospheric applications: The importance of appropriate weighting

Cheng Wu[1,2] and Jian Zhen Yu[3,4,5]

[1]Institute of Mass Spectrometer and Atmospheric Environment, Jinan University, Guangzhou 510632, China

[2]Guangdong Provincial Engineering Research Center for on-line source apportionment system of air pollution, Guangzhou 510632, China

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

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

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

Corresponding to: Cheng Wu (wucheng.vip@foxmail.com) and Jian Zhen Yu (jian.yu@ust.hk)

Abstract

Linear regression techniques are widely used in atmospheric science, but are often improperly applied due to lack of consideration or inappropriate handling of measurement uncertainty. In this work, numerical experiments are performed to evaluate the performance of five linear regression techniques, significantly extending previous works by Chu and Saylor. The tested are Ordinary Least Square (OLS), Deming Regression (DR), Orthogonal Distance Regression (ODR), Weighted ODR (WODR), and York regression (YR). We first introduce a new data generation scheme that employs the Mersenne Twister (MT) pseudorandom number generator. The numerical simulations are also improved by: (a) refining the parameterization of non-linear measurement uncertainties, (b) inclusion of a linear measurement uncertainty, (c) inclusion of WODR for comparison. Results show that DR, WODR and YR produce an accurate slope, but the intercept by WODR and YR is overestimated and the degree of bias is more pronounced with a low R^2 XY dataset. The importance of a properly weighting parameter λ in DR is investigated by sensitivity tests, and it is found an improper λ in DR can leads to a bias in both the slope and intercept estimation. Because the λ calculation depends on the actual form of the measurement error, it is essential to determine the exact form of measurement error in the XY data during the measurement stage. If discrepancy exist between measurement error of data and measurement uncertainty used for regression, DR, WODR and YR can provide the least biases in slope and intercept among all tested regression techniques. For these reasons, DR, WODR and YR are recommended for atmospheric studies when both X and Y data have measurement errors.

1 Introduction

Linear regression is heavily used in atmospheric science to derive the slope and intercept of XY datasets. Examples of linear regression applications include primary OC (organic carbon) and EC (elemental carbon) ratio estimation (Turpin and Huntzicker, 1995), MAE (mass absorption efficiency) estimation from light absorption and EC mass (Moosmüller et al., 1998), source apportionment of polycyclic aromatic hydrocarbons using CO and NO_x as combustion tracers (Lim et al., 1999), gas-phase reaction rate determination (Brauers and Finlayson-Pitts, 1997), inter-instrument comparison (Bauer et al., 2009; Cross et al., 2010; von Bobrutzki et al., 2010; Zieger et al., 2011; Huang et al., 2014; Zhou et al., 2016), light extinction budget reconstruction (Malm et al., 1994; Watson, 2002), comparison between modeling and measurement (Petäjä et al., 2009), emission factor study (Janhäll et al., 2010), retrieval of shortwave cloud forcing (Cess et al., 1995), calculation of pollutant growth rate (Richter et al., 2005), estimation of ground level $PM_{2.5}$ from MODIS data (Wang and Christopher, 2003), distinguishing OC origin from biomass burning using K^+ as a tracer (Duan et al., 2004) and emission type identification by the EC/CO ratio (Chen et al., 2001).

Ordinary least squares (OLS) regression is the most widely used method due to its simplicity. In OLS, it is assumed that independent variables are error free. This is the case for certain applications, such as determining a calibration curve of an instrument in analytical chemistry. For example, a known amount of analyte (e.g., through weighing) can be used to calibrate the instrument output response (e.g., voltage). However, in many other applications, such as inter-instrument comparison, X and Y (from two instruments) may have comparable degrees of uncertainty. This deviation from the underlying assumption in OLS would produce biased slope and intercept when OLS is applied to the dataset.

To overcome the drawback of OLS, a number of error-in-variable regression models (also known as bivariate fittings (Cantrell, 2008) or total least-squares methods (Markovsky and Van Huffel, 2007) arise. Deming (1943) proposed an approach by minimizing sum of squares of X and Y residuals. A closed-form solution of Deming regression (DR) was provided by York (1966). Method comparison work of various regression techniques by Cornbleet and Gochman (1979) found significant error in OLS slope estimation when the relative standard deviation (RSD) of measurement error in "X" exceeded 20%, while DR was found to reach a more accurate slope estimation. In an early application of the EC tracer method, Turpin and Huntzicker (1995) realized the limitation of OLS since OC and EC have comparable measurement uncertainty, thus recommended the use of DR for $(OC/EC)_{pri}$ (primary OC to EC ratio) estimation. Ayers (2001) conducted a simple numerical experiment and concluded that reduced major axis regression (RMA) is more suitable for air quality data regression analysis. Linnet (1999) pointed out that when applying DR for inter-method (or inter-instrument) comparison, special attention should be paid to the sample size. If the range ratio (max/min) is relatively small (e.g., less than 2), more samples are needed to obtain statistically significant results.

In principle, a best-fit regression line should have greater dependence on the more precise data points rather than the less reliable ones. Chu (2005) performed a comparison study of OLS and DR specifically focusing on the EC tracer method application, and found the slope estimated by DR is closer to the correct value than OLS but may still overestimate the ideal value. Saylor et al. (2006) extended the comparison work of Chu (2005) by including a regression technique developed by York et al. (2004). They found that the slope overestimation by DR in the study of Chu (2005) was due to improper configuration of the weighting parameter, λ. This λ value is the key to handling the uneven errors between data points for the best-fit line calculation. This example demonstrates the importance of appropriate weighting in the calculation of best-bit line for error-in-variable regression model, which is overlooked in many studies.

In this study, we extend the work by Saylor et al. (2006) to achieve four objectives. The first is to propose a new data generation scheme by applying the Mersenne Twister (MT) pseudorandom number generator for evaluation of linear regression techniques. In the study of Chu (2005), data generation is achieved by a varietal sine function, which has limitations in sample size, sample distribution, and nonadjustable correlation (R^2) between X and Y. In comparison, the MT data generation provides more flexibility, permitting adjustable sample size, XY correlation and distribution. The second is to develop a non-linear measurement error parameterization scheme for use in the regression method. The third is to incorporate linear measurement errors in the regression methods. In the work by Chu (2005) and Saylor et al. (2006), the relative measurement uncertainty (γ_{Unc}) is non-linear with concentration, but a constant γ_{Unc} is often applied on atmospheric instruments due to its simplicity. The fourth is to include weighted orthogonal distance regression (WODR) for comparison. Abbreviations and symbols used in this study are summarized in Table 1 for quick lookup.

2 Description of regression techniques compared in this study

Ordinary least squares (OLS) method. OLS only considers the errors in dependent variables (Y). OLS regression is achieved by minimizing the sum of squares (S) in the Y residuals:

$$S = \sum_{i=1}^{n}(y_i - Y_i)^2 \quad (1)$$

where Y_i are observed Y data points while y_i are regressed Y data points of the regression line.

Orthogonal distance regression (ODR). ODR minimizes the sum of the squared orthogonal distances from all data points to the regressed line and considers equal error variances:

$$S = \sum_{i=1}^{n}[(x_i - X_i)^2 + (y_i - Y_i)^2] \quad (2)$$

Weighted orthogonal distance regression (WODR). Unlike ODR that considers even error in X and Y, weightings based on measurement errors in both X and Y are considered in WODR when minimizing the sum of squared orthogonal distance from the data points to the regression line (Carroll and Ruppert, 1996):

$$S = \sum_{i=1}^{n}[(x_i - X_i)^2 + (y_i - Y_i)^2/\eta] \quad (3)$$

where η is error variance ratio. Implementation of ODR and WODR in Igor was done by the computer routine ODRPACK95 (Boggs et al., 1989; Zwolak et al., 2007).

Deming regression (DR). Deming (1943) proposed the following function to minimize both the X and Y residuals,

$$S = \sum_{i=1}^{n}[\omega(X_i)(x_i - X_i)^2 + \omega(Y_i)(y_i - Y_i)^2] \quad (4)$$

where X_i and Y_i are observed data points and x_i and y_i are regressed data points. Individual data points are weighted based on errors in X_i and Y_i,

$$\omega(X_i) = \frac{1}{{\sigma_{X_i}}^2}, \quad \omega(Y_i) = \frac{1}{{\sigma_{Y_i}}^2} \tag{5}$$

where σ_{X_i} and σ_{Y_i} are the standard deviation of the error in measurement of X_i and Y_i respectively. The closed form solutions for slope and intercept of DR are shown in Appendix A.

York regression (YR). The York method (York et al., 2004) introduces the correlation coefficient of errors in X and Y into the minimization function.

$$S = \sum_{i=1}^{n}\left[\omega(X_i)(x_i - X_i)^2 - 2r_i\sqrt{\omega(X_i)\omega(Y_i)}(x_i - X_i)(y_i - Y_i) + \omega(Y_i)(y_i - Y_i)^2\right]\frac{1}{1-{r_i}^2} \tag{6}$$

where r_i is the correlation coefficient between measurement errors in X_i and Y_i. The slope and intercept of YR are calculated iteratively through the formulas in Appendix A.

3 Data description

Two types of data are used for regression comparison. The first type is synthetic data generated by computer programs, which can be used in the EC tracer method (Turpin and Huntzicker, 1995) to demonstrate the regression application. The true "slope" and "intercept" are assigned during data generation, allowing quantitative comparison of the bias of each regression scheme. The second type of data comes from ambient measurement of light absorption, OC and EC in Guangzhou for demonstration in a real-world application.

3.1 Synthetic XY data generation

In this study, numerical simulations are conducted in Igor Pro (WaveMetrics, Inc. Lake Oswego, OR, USA) through custom codes. Two types of generation schemes are employed, one is based on the MT pseudorandom number generator (Matsumoto and Nishimura, 1998) and the other is based on the sine function described by Chu (2005).

The general form of linear regression on XY data can be written as:

$$\mathrm{Y} = \mathrm{kX} + \mathrm{b} \tag{7}$$

Here k is the regressed slope and b is the intercept. The underlying meaning is that, Y can be decomposed into two parts. One part is correlated with X, and the ratio is defined by k. The other part of Y is constant and independent of X and regarded as b.

To make the discussion easier to follow, we intentionally avoid discussion using the abstract general form and instead opt to use a real-world application case in atmospheric science. Linear regression had been heavily applied on OC and EC data, here we use OC and EC data as an example to demonstrate the regression application in atmospheric science. In the EC tracer method, OC (mixture) is Y and EC (tracer) is X. OC can be decomposed into three components based on their formation pathway:

$$OC = POC_{comb} + POC_{non-comb} + SOC \quad (8)$$

Here POC_{comb} is primary OC from combustion. $POC_{non-comb}$ is primary OC emitted from non-combustion activities. SOC is secondary OC formed during atmospheric aging. Since POC_{comb} is co-emitted with EC and well correlated with each other, their relationship can be parameterized as:

$$POC_{comb} = (OC/EC)_{pri} \times EC \quad (9)$$

By carefully selecting an OC and EC subset when SOC is very low (considered as approximately zero), the combination of Eqs. (8) & (9) become:

$$POC = (OC/EC)_{pri} \times EC + POC_{non-comb} \quad (10)$$

The regressed slope of POC (Y) against EC (X) represents $(OC/EC)_{pri}$ (k in Eq.(7)). The regressed intercept become $POC_{non-comb}$ (b in Eq. (7)). With known $(OC/EC)_{pri}$ and $POC_{non-comb}$, SOC can be estimated by:

$$SOC = OC - ((OC/EC)_{pri} \times EC + POC_{non-comb}) \quad (11)$$

The data generation starts from EC (X values). Once EC is generated, POC_{comb} (the part of Y that is correlated with X) can be obtained by multiplying EC with a preset constant, $(OC/EC)_{pri}$ (slope k). Then the other preset constant $POC_{non-comb}$ is added to POC_{comb} and the sum becomes POC (Y values). To simulate the real-world situation, measurement errors are added on X and Y values. Details of synthesized measurement error are discussed in the next section. Implementation of data generation by two types of mathematical schemes are explained in section 3.1.2 and 3.1.3 respectively.

3.1.1 Parameterization of synthesized measurement uncertainty

Weighting of variables is a crucial input for errors-in-variables linear regression methods such as DR, YR and WODR. In practice, the weights are usually defined as the inverse of the measurement error variance (Eq. (5)). When measurement errors are considered, measured concentrations ($Conc._{measured}$) are simulated by adding measurement uncertainties ($\varepsilon_{Conc.}$) to the true concentrations ($Conc._{true}$):

$$Conc._{measured} = Conc._{true} + \varepsilon_{Conc.} \tag{12}$$

Here $\varepsilon_{Conc.}$ is the random error following an even distribution with an average of 0, the range of which is constrained by:

$$-\gamma_{Unc} \times Conc._{true} \leq \varepsilon_{Conc.} \leq +\gamma_{Unc} \times Conc._{true} \tag{13}$$

The γ_{Unc} is a dimensionless factor that describes the fractional measurement uncertainties relative to the true concentration ($Conc._{true}$). γ_{Unc} could be a function of $Conc._{true}$ (Thompson, 1988) or a constant. The term $\gamma_{Unc} \times Conc._{true}$ defines the boundary of random measurement errors.

Two types of measurement error are considered in this study. The first type is $\gamma_{Unc-nonlinear}$. In the data generation scheme of Chu (2005) for the measurement uncertainties (ε_{POC} and ε_{EC}), $\gamma_{Unc-nonlinear}$ is non-linearly related to $Conc._{true}$:

$$\gamma_{Unc-nonlinear} = \frac{1}{\sqrt{Conc._{true}}} \tag{14}$$

then Eq. (13) for POC and EC become:

$$-\frac{1}{\sqrt{POC_{true}}} \times POC_{true} \leq \varepsilon_{POC} \leq +\frac{1}{\sqrt{POC_{true}}} \times POC_{true} \tag{15}$$

$$-\frac{1}{\sqrt{EC_{true}}} \times EC_{true} \leq \varepsilon_{EC} \leq +\frac{1}{\sqrt{EC_{true}}} \times EC_{true} \tag{16}$$

In Eq. (14), the γ_{Unc} decreases as concentration increases, since low concentrations are usually more challenging to measure. As a result, the $\gamma_{Unc-nonlinear}$ defined in Eq. (14) is more realistic than the constant approach, but there are two limitations. First, the physical meaning of the uncertainty unit is lost. If the unit of OC is μg m^{-3}, then the unit of ε_{OC} becomes $\sqrt{\mu g\ m^{-3}}$. Second, the concentration is not normalized by a consistent relative value, making it sensitive to the X and Y units used. For example, if

POC_{true}=0.9 μg m^{-3}, then ε_{POC}= ±0.95 μg m^{-3} and γ_{Unc} = 105%, but by changing the concentration unit to POC_{true}=900 ng m^{-3}, then ε_{OC}= ±30 ng m^{-3} and γ_{Unc} = 3%. To overcome these deficiencies, we propose to modify Eq. (14) to:

$$\gamma_{Unc} = \sqrt{\frac{LOD}{Conc._{true}}} \times \alpha \tag{17}$$

here LOD (limit of detection) is introduced to generate a dimensionless γ_{Unc}. α is a dimensionless adjustable factor to control the position of γ_{Unc} curve on the concentration axis, which is indicated by the value of γ_{Unc} at LOD level. As shown in Figure 1a, at different values of α (α =1, 0.5 and 0.3), the corresponding γ_{Unc} at the same LOD level would be 100%, 50% and 30% respectively. By changing α, the location of the γ_{Unc} curve on X axis direction can be set, using the γ_{Unc} at LOD as the reference point. Then Eq. (17) for POC and EC become:

$$-\sqrt{\frac{LOD_{POC}}{POC_{true}}} \times \alpha_{POC} \times POC_{true} \leq \varepsilon_{POC} \leq +\sqrt{\frac{LOD_{POC}}{POC_{true}}} \times \alpha_{POC} \times POC_{true} \tag{18}$$

$$-\sqrt{\frac{LOD_{EC}}{EC_{true}}} \times \alpha_{EC} \times EC_{true} \leq \varepsilon_{EC} \leq +\sqrt{\frac{LOD_{EC}}{EC_{true}}} \times \alpha_{EC} \times EC_{true} \tag{19}$$

With the modified $\gamma_{Unc-nonlinear}$ parameterization, concentrations of POC and EC are normalized by a corresponding LOD, which maintains unit consistency between POC_{true} and ε_{POC} and EC_{true} and ε_{EC}, and eliminates dependency on the concentration unit.

Uniform distribution had been used in previous studies (Cox et al., 2003; Chu, 2005; Saylor et al., 2006) and is adopted in this study to parameterize measurement error. For a uniform distribution in the interval [a,b], the variance is $\frac{1}{12}(a-b)^2$. Since ε_{POC} and ε_{EC} follows a uniform distribution in the interval as given by Eqs. (18) and (19), the weights in DR and YR (inverse of variance) become:

$$\omega(X_i) = \frac{1}{{\sigma_{X_i}}^2} = \frac{3}{EC_{true} \times LOD_{EC} \times {\alpha_{EC}}^2} \tag{20}$$

$$\omega(Y_i) = \frac{1}{{\sigma_{Y_i}}^2} = \frac{3}{POC_{true} \times LOD_{POC} \times {\alpha_{POC}}^2} \tag{21}$$

The parameter λ in Deming regression is then determined:

$$\lambda = \frac{\omega(X_i)}{\omega(Y_i)} = \frac{POC_{true} \times LOD_{POC} \times \alpha_{POC}{}^2}{EC_{true} \times LOD_{EC} \times \alpha_{EC}{}^2} \quad (22)$$

Besides the $\gamma_{Unc-nonlinear}$ discussed above, a second type measurement uncertainty parameterized by a constant proportional factor, $\gamma_{Unc-linear}$, is very common in atmospheric applications:

$$-\gamma_{POCunc} \times POC_{true} \leq \varepsilon_{POC} \leq +\gamma_{POCunc} \times POC_{true} \quad (23)$$

$$-\gamma_{ECunc} \times EC_{true} \leq \varepsilon_{EC} \leq +\gamma_{ECunc} \times EC_{true} \quad (24)$$

where γ_{POCunc} and γ_{ECunc} are the relative measurement uncertainties, e.g., for relative measurement uncertainty of 10%, γ_{Unc}=0.1. As a result, the measurement error is linearly proportional to the concentration. An example comparison of $\gamma_{Unc-nonlinear}$ and $\gamma_{Unc-linear}$ is shown in Figure 1b. For $\gamma_{Unc-linear}$, the weights become:

$$\omega(X_i) = \frac{1}{\sigma_{X_i}{}^2} = \frac{3}{(\gamma_{ECunc} \times EC_{true})^2} \quad (25)$$

$$\omega(Y_i) = \frac{1}{\sigma_{Y_i}{}^2} = \frac{3}{(\gamma_{POCunc} \times POC_{true})^2} \quad (26)$$

and λ for Deming regression can be determined:

$$\lambda = \frac{\omega(X_i)}{\omega(Y_i)} = \frac{(\gamma_{POCunc} \times POC_{true})^2}{(\gamma_{ECunc} \times EC_{true})^2} \quad (27)$$

3.1.2 XY data generation by Mersenne Twister (MT) generator following a specific distribution

The Mersenne twister (MT) is a pseudorandom number generator (PRNG) developed by Matsumoto and Nishimura (1998). MT has been widely adopted by mainstream numerical analysis software (e.g., Matlab, SPSS, SAS and Igor Pro) as well as popular programing languages (e.g., R, Python, IDL, C++ and PHP). Data generation using MT provides a few advantages: (1) Frequency distribution can be easily assigned during the data generation process, allowing straightforward simulation of the frequency distribution characteristics (e.g., Gaussian or Log-normal) observed in ambient measurements; (2) The inputs for data generation are simply the mean and standard deviation of the data series and can be changed easily by the user; (3) The correlation (R^2) between X and Y can be manipulated easily during the data generation to satisfy various purposes; (4) Unlike the sine function described by Chu (2005) that has a sample size limitation of 120, the sample size in MT data generation is highly flexible.

In this section, we will use POC as Y and EC as X as an example to explain the data generation. Procedure of applying MT to simulate ambient POC and EC data can be found in our previous study (Wu and Yu, 2016). Details of the data generation steps are shown in Figure 2 and described below. The first step is generation of EC_{true} by MT. In our previous study, it was found that ambient POC and EC data follow a lognormal distribution in various locations of the Pearl River Delta (PRD) region. Therefore, lognormal distributions are adopted during EC_{true} generation. A range of average concentration and relative standard deviation (RSD) from ambient samples are considered in formulating the lognormal distribution. The second step is to generate POC_{comb}. As shown in Figure 2, POC_{comb} is generated by multiplying EC_{true} with $(OC/EC)_{pri}$. Instead of having a Gaussian distribution, $(OC/EC)_{pri}$ in this study is a single value, which favors direct comparison between the true value of $(OC/EC)_{pri}$ and $(OC/EC)_{pri}$ estimated from the regression slope. The third step is generation of POC_{true} by adding $POC_{non\text{-}comb}$ onto POC_{comb}. Instead of having a distribution, $POC_{non\text{-}comb}$ in this study is a single value, which favors direct comparison between the true value of $POC_{non\text{-}comb}$ and $POC_{non\text{-}comb}$ estimated from the regression intercept. The fourth step is to compute ε_{POC} and ε_{EC}. As discussed in section 3.1, two types of measurement errors are considered for ε_{POC} and ε_{EC} calculation: $\gamma_{Unc-nonlinear}$ and $\gamma_{Unc-linear}$. In the last step, $POC_{measured}$ and $EC_{measured}$ are calculated following Eq. (12), i.e., applying measurement errors on POC_{true} and EC_{true}. Then $POC_{measured}$ and $EC_{measured}$ can be used as Y and X respectively to test the performance of various regression techniques. An Igor Pro based program with graphical user interface (GUI) is developed to facilitate the MT data generation for OC and EC. A brief introduction is given in the Supplemental Information.

3.1.3 XY data generation by the sine function of Chu (2005)

Beside MT, the inclusion of the sine function data generation schemes in this study mainly serves two purposes. First, the sine function scheme had been adopted by two previous studies (Chu, 2005; Saylor et al., 2006), the inclusion of this scheme can help to verify whether the codes in Igor for various regression approaches can yield the same results from the two previous studies. Second, crosscheck between results from sine function and MT can provides circumstantial evidence that the MT scheme works as expected.

In this section, XY data generation by sine functions is demonstrated using POC as Y and EC as X. There are four steps in POC and EC data generation as shown by the flowchart in Figure S1. Details are explained as follows: (1) The first step is to generate POC and EC (Chu, 2005):

$$POC_{comb} = 14 + 12(\sin(\frac{x}{\tau}) + \sin(x - \phi)) \quad (28)$$

$$EC_{true} = 3.5 + 3(\sin(\frac{x}{\tau}) + \sin(x - \phi)) \quad (29)$$

Here x is the elapsed hour (x=1,2,3……n; n≤120), τ is used to adjust the width of each peak, and ϕ is used to adjust the phase of the sine wave. The constants 14 and 3.5 are used to lift the sine wave to the positive range of the Y axis. An example of data generation by the sine functions of Chu (2005) is shown in Figure 3. Dividing Eq. (28) by Eq. (29) yields a value of 4. In this way the exact relation between POC and EC is defined clearly as $(OC/EC)_{pri}$ = 4. (2) With POC_{comb} and EC_{true} generated, the second step is to add $POC_{non\text{-}comb}$ to POC_{comb} to compute POC_{true}. As for $POC_{non\text{-}comb}$, a single value is assigned and added to all POC following Eq. (10). Then the goodness of the regression intercept can be evaluated by comparing the regressed intercept with preset $POC_{non\text{-}comb}$. (3) The third step is to compute ε_{POC} and ε_{EC}, considering both $\gamma_{Unc-nonlinear}$ and $\gamma_{Unc-linear}$. (4) The last step is to apply measurement errors on POC_{true} and EC_{true} following Eq. (12). Then $POC_{measured}$ and $EC_{measured}$ can be used as Y and X respectively to evaluate the performance of various regression techniques.

3.2 Ambient measurement of σ_{abs} and EC

Sampling was conducted from Feb 2012 to Jan 2013 at the suburban Nancun (NC) site (23° 0'11.82"N, 113°21'18.04"E), which is situated on the top of the highest peak (141 m ASL) in the Panyu district of Guangzhou. This site is located at the geographic center of Pearl River Delta region (PRD), making it a good location for representing the average atmospheric mixing characteristics of city clusters in the PRD region. Light absorption measurements were performed by a 7-λ Aethalometer (AE-31, Magee

Scientific Company, Berkeley, CA, USA). EC mass concentrations were measured by a real time ECOC analyzer (Model RT-4, Sunset Laboratory Inc., Tigard, Oregon, USA). Both instruments utilized inlets with a 2.5 μm particle diameter cutoff. The algorithm of Weingartner et al. (2003) was adopted to correct the sampling artifacts (aerosol loading, filter matrix and scattering effect) (Coen et al., 2010) root in Aethalometer measurement. A customized computor program with graphical user interface, Aethalometer data processor (Wu et al., 2017), was developed to perform the data correction and detailed descriptions can be found in https://sites.google.com/site/wuchengust. More details of the measurements can be found in Wu et al. (2017).

4 Comparison study using synthetic data

In the following comparisons, six regression approaches are compared using two data generation schemes (Chu sine function and MT) separately, as illustrated in Figure 4. Each data generation scheme considers both $\gamma_{Unc-nonlinear}$ and $\gamma_{Unc-linear}$ in measurement error parameterization. In total, 18 cases are tested with different combination of data generation schemes, measurement error parameterization schemes, true slope and intercept settings. For each case, six regression approaches are tested, including OLS, DR ($\lambda = 1$), DR ($\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$), ODR, WODR and YR. In commercial software (e.g., Origin, SigmaPlot, GraphPad Prism, etc), λ in DR is set to 1 by default if not specified. As indicated by Saylor et al. (2006), the bias observed in the study of Chu (2005) is likely due to $\lambda = 1$ in DR. The purpose of including DR ($\lambda = 1$) in this study is to examine the potential bias using the default input in many software products. The six regression approaches are considered to examine the sensitivity of regression results to various parameters used in data generation. For each case, 5000 runs are performed to obtain statistically significant results, as recommended by Saylor et al. (2006). The mean slope and intercept from 5000 runs is compared with the true value assigned during data generation. If the difference is <5%, the result is considered unbiased.

4.1 Comparison results using the data set of Chu (2005)

In this section, the scheme of Chu (2005) is adopted for data generation to obtain a benchmark of six regression approaches. With different setup of slope, intercept and γ_{Unc} , 6 cases (Case 1 ~ 6) are studied and the results are discussed below.

4.1.1 Results with $\gamma_{Unc-nonlinear}$

A comparison of the regression techniques results with $\gamma_{Unc-nonlinear}$ (following Eqs. (18) & (19)) are summarized in Table 2. LOD_{POC} , LOD_{EC}, α_{POC} and α_{EC} are all set to 1 to reproduce the data studied by Chu (2005) and Saylor et al. (2006). Two sets of true slope and intercept are considered (Case 1: Slope=4, Intercept=0; Case 2: Slope=4, Intercept=3) to examine if any results are sensitive to the non-zero intercept. The R^2 (POC, EC) from 5000 runs for both case 1 and 2 are 0.67±0.03.

As shown in Figure 5, for the zero-intercept case (Case 1), OLS significantly underestimates the slope (2.95±0.14) while overestimates the intercept (5.84±0.78). This result indicates that OLS is not suitable for errors-in-variables linear regression, consistent with similar analysis results from Chu (2005) and Saylor et al. (2006). With DR, if the λ is properly calculated by weights ($\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$), unbiased slope (4.01±0.25) and intercept (-0.04±1.28) are obtained, however, results from DR with λ=1 shows obvious bias in the slope (4.27±0.27) and intercept (-1.45±1.36). ODR also produces biased slope (4.27±0.27) and intercept (-1.45±1.36), which are identical to results of DR when λ=1. With WODR, unbiased slope (3.98±0.22) is observed, but the intercept is overestimated (1.12±1.02). Results of YR are identical to WODR. For Case 2 (slope=4, intercept=3), slopes from all six regression approaches are consistent with Case 1 (Table 2). The Case 2 intercepts are equal to the Case 1 intercepts plus 3, implying that all the regression methods are not sensitive to a non-zero intercept.

For case 3, LOD_{POC}=0.5, LOD_{EC}=0.5, α_{POC}=0.5, α_{EC}=0.5 are adopted (Table 2), leading to an offset to the left of $\gamma_{Unc-nonlinear}$ (blue curve) compared to Case 1 and 2 (black curve) in Figure 1. As a result, for the same concentration of EC and OC in Case 3, the $\gamma_{Unc-nonlinear}$ is smaller than in Case 1 and Case 2 as indicated by higher the R^2 (0.95±0.01 for Case 3, Table 2). With a smaller measurement uncertainty, the degree of bias in Case 3 is smaller than Case 1. For example, OLS slope is less biased in Case

3 (3.83±0.08) compare to Case 1 (2.94±0.14). Similarly, the slope (4.03±0.09) and intercept (-0.18±0.44) of DR (λ=1) exhibit a much smaller bias with a smaller measurement uncertainty, implying that the degree of bias by improperly weighting in DR, WODR and YR is associated with the degree of measurement uncertainty. A higher measurement uncertainty results in larger bias in slope and intercept.

An uneven LOD_{POC} and LOD_{EC} is tested in Case 4 with LOD_{POC}=1, LOD_{EC}=0.5, α_{POC}=0.5, α_{EC}=0.5, which yield a R^2(POC, EC) of 0.78±0.02. The results are similar to Case 1. For DR ($\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$) unbiased slope and intercept are obtained. For WODR and YR, unbiased slopes are reported with a small bias in the intercepts. Large bias values are observed in both the slopes and intercepts in Case 4 using OLS, DR ($\lambda = 1$) and ODR.

4.1.2 Results with $\gamma_{Unc-linear}$

Cases 5 and 6 represent the results from using $\gamma_{Unc-linear}$ and are shown in Table 2. γ_{Unc} is set to be 30% to achieve a R^2 (POC, EC) of 0.7, a value close to the R^2 in studies of Chu (2005) and Saylor et al. (2006). In Case 5 (slope=4, intercept=0), unbiased slopes and intercepts are determined by DR ($\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$), WODR and YR. OLS underestimates the slope (3.32 ±0.20) and overestimates intercept (3.77 ±0.90), while DR ($\lambda = 1$) and ODR overestimate the slopes (4.75 ±0.30) and underestimates the intercepts (-4.14 ±1.36). In Case 6 (slope=4, intercept=3), results similar to Case 5 are obtained. It is worth noting that although the mean intercept (3.05±1.22) of DR ($\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$), is closest to the true value (intercept=3), the deviations are much larger than for WODR (2.72±0.74).

4.2 Comparison results using data generated by MT

In this section, MT is adopted for data generation to obtain a benchmark of six regression approaches. Both $\gamma_{Unc-nonlinear}$ and $\gamma_{Unc-linear}$ are considered. With different configuration of slope, intercept and γ_{Unc}, 12 cases (Case 7 ~ Case 18) are studied and the results are discussed below.

4.2.1 $\gamma_{Unc-nonlinear}$ results

Cases 7 and 8 use data generated by MT and $\gamma_{Unc-nonlinear}$ with results shown in Table 2. In Case 7 (slope=4, intercept=0, LOD_{POC}=1, LOD_{EC}=1, α_{POC}=1, α_{EC}=1), unbiased slope (4.00 ±0.03) and intercept (0.00 ±0.17) is estimated by DR ($\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$). WODR and YR yield unbiased slopes (3.96 ±0.03) but overestimate the intercepts (1.21 ±0.13). DR ($\lambda = 1$) and ODR report slightly biased slopes (4.17 ±0.04) with biased intercepts (-0.94 ±0.18). OLS underestimates the slope (3.22 ±0.03) and overestimates the intercept (4.30 ±0.14). In Case 8 (slope=4, intercept=3, LOD_{POC}=1, LOD_{EC}=1, α_{POC}=1, α_{EC}=1), DR ($\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$) provides unbiased slope (4.00 ±0.03) and intercept (3.00 ±0.18) estimations. WODR and YR report unbiased slopes (3.97 ±0.03) and overestimate intercepts (4.11 ±0.13). OLS, DR ($\lambda = 1$) and ODR report biased slopes and intercepts.

To test the overestimation/underestimation dependency on the true slope, Case 9 (slope=0.5, intercept=0, LOD_{POC} =1, LOD_{EC} =1, α_{POC} =1, α_{EC} =1) and case 10 (slope=0.5, intercept=3, LOD_{POC}=1, LOD_{EC}=1, α_{POC}=1, α_{EC}=1) are conducted and the results are shown in Table 2. Unlike the overestimation observed in Case 1~Case 8, DR ($\lambda = 1$) and ODR underestimate the slopes (0.46 ±0.01) in Case 9. In case 10, DR ($\lambda = 1$), DR ($\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$) and ODR report unbiased slopes and intercepts. Case 11 and case 12 test the bias when the true slope is 1 as shown in Table 2. In Case 11 (intercept=0), all regression approaches except OLS can provide unbiased results. In Case 12, all regression approaches report unbiased slopes except OLS, but DR ($\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$) is the only regression approach that report unbiased intercept.

These results imply that if the true slope is less than 1, the improper weighting ($\lambda = 1$) in Deming regression and ODR without weighting tends to underestimate slope. If the true slope is 1, these two estimators can provide unbiased results. If the true slope is larger than 1, the improper weighting ($\lambda = 1$) in Deming regression and ODR without weighting tends to overestimate slope.

4.2.2 $\gamma_{Unc-linear}$ results

Cases 13 and 14 (Table 2) represent the results from using $\gamma_{Unc-linear}$ (30%) and data generated from MT. For case 13 (slope=4, intercept=0), DR ($\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$), WODR and YR provide the best estimation of slopes and intercepts. DR ($\lambda = 1$) and ODR overestimate slopes (4.53 ±0.05) and underestimate intercepts (-2.94 ±0.24). For case 14 (slope=4, intercept=3), DR ($\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$), WODR and YR provide an unbiased estimation of slopes. But DR ($\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$) is the only regression approaches reports unbiased intercept (3.08 ±0.23). Cases 15 and 16 are tested to investigate whether the results are different if the true slope is smaller than 1. As shown in Table 2, the results are similar to case 13&14 that DR ($\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$) can provide unbiased slope and intercept while WODR and YR can provide unbiased slopes but biased intercepts. Cases 17 and 18 are tested to see if the results are the same for a special case when the true slope is 1. As shown in Table 2, the results are similar to case 13&14, implying that these results are not sensitive to the special case when the true slope is 1.

4.3 The importance of appropriate λ input for Deming regression

As discussed above, inappropriate λ assignment in the Deming regression (e.g., λ=1 by default for many commercial software) leads to biased slope and intercept. Beside λ =1, inappropriate λ input due to improper handling of measurement uncertainty can also result in bias for Deming regression. An example is shown in Figure S2. Data is generated by MT with following parameters: slope=4, intercept=0, and $\gamma_{Unc-linear}$ (30%). Figure S2 a&b demonstrates that when an appropriate λ is provided (following $\gamma_{Unc-linear}$, $\lambda = \frac{POC^2}{EC^2}$), unbiased slopes and intercepts are obtained. If an improper λ is used due to a mismatched measurement uncertainty assumption ($\gamma_{Unc-nonlinear}$, $\lambda = \frac{POC}{EC}$), the slopes are overestimated (Figure S2c, 4.37±0.05) and intercepts are underestimated (Figure S2d，-2.01±0.24). This result emphasizes the importance of determining the correct form of measurement uncertainty in ambient samples, since λ is a crucial parameter in Deming regression.

In the λ calculation, different representations for POC and EC, including mean, median and mode, are tested as shown in Figure S3. The results show that when X and Y have a similar distribution (e.g., both are log-normal), any of mean, median or mode can be used for the λ calculation.

4.4 Caveats of regressions with unknown X and Y uncertainties

When applying linear regression on real world data, it happens that a priori error in one of the variables is unknown, or the measurement error described cannot be trusted. In other words, that would be certain degree of discrepancy between the measurement error used for linear regression and measurement error embed in the data. It is common that measurement error cannot be determined due to the lack of duplicated or collocated measurements and an arbitrarily assumed uncertainty is used. For example, Flanagan et al. (2006) found that the whole-system uncertainty retrieved by data from collocated sampler is different from the arbitrarily assumed 5% uncertainty, which is previously used by the Speciation Trends Network (STN). In addition, the degree of discrepancy between the actual uncertainty by collocated samples and arbitrarily assumed uncertainty also varied by different chemical species. To investigate the impact of such cases on different regression approaches, two tests are conducted. In Test A, the actual measurement error for X is fixed at 30% while γ_{Unc} for Y varied from 1% to 50%. The assumed measurement error for regression is 10% for both X and Y. Results of Test A are shown in Figure 6 a&b. For OLS, the slopes are underestimated (-14 ~ -12%) and intercepts are overestimated (90 ~ 103%). The biases in OLS slope and intercept are independent of variations in γ_{Unc_Y}. ODR and DR ($\lambda = 1$) yield similar results with overestimated slopes (0 ~ 44%) and underestimated intercepts (-330 ~ 0%). The degree of bias in slopes and intercepts depends on γ_{Unc_Y}. WODR, DR ($\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$) and YR performed much better than other regression approaches in Test A, with a smaller bias in both slopes (-8 ~ 12%) and intercepts -98 ~ 55%).

The results of Test B are shown in Figure 6 c&d. which has a fixed γ_{Unc_Y} of 30% and γ_{Unc_X} varied between 1 ~ 50%. The assumed measurement error for regression is 10% for both X and Y. OLS underestimates slopes (-29 ~-0.2%) and overestimates intercepts (2 ~ 209%) in Test B. In contrast to Test A which slope and intercept biases are independent of variations in γ_{Unc_Y}, the OLS slope and intercept biases in Test B exhibit dependency on γ_{Unc_X}. The reason behind is because OLS only considers errors in Y, while X is assumed to be error free. ODR and DR ($\lambda = 1$) yield similar results with overestimated slopes (11 ~ 18%) and underestimated intercepts (-144 ~ -87%). The degree of biases in slopes and intercepts is relatively independent to the γ_{Unc_X}. WODR, DR ($\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$) and YR performed much better than other regression approaches in Test B, with a smaller bias in both slopes (-14 ~ 8%) and intercepts (-59 ~ 106%).

The results from these two tests suggest that, in case of one of the measurement error described cannot be trusted or a priori error in one of the variables is unknown, WODR, DR ($\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$) and YR should be used instead of ODR, DR ($\lambda = 1$) and OLS. This conclusion also agrees with section 4.1 and 4.2. The results also suggest that, in general, the magnitude of bias in slope estimation by these regression approaches are smaller than those for intercept. In other words, slope is a more reliable quantity compare to intercept when extracting quantitative information from linear regressions.

5 Regression applications to ambient data

This section demonstrates the application of the 6 regression approaches on a light absorption coefficient and EC dataset collected in a suburban site in Guangzhou. As mentioned in the last section, measurement uncertainties are crucial inputs for DR, YR and WODR. The measurement precision of Aethalometer is 5% (Hansen, 2005) while EC by RT-ECOC analyzer is 24% (Bauer et al., 2009). These measurement uncertainties are used in DR, YR and WODR calculation. The data-set contains 6926 data points with a R^2 of 0.92.

As shown in Figure 7, Y axis is light absorption at 520 nm (σ_{abs520}) and the X axis is EC mass concentration. The regressed slopes represent the mass absorption efficiency (MAE) of EC at 520 nm, ranging from 13.66 to 15.94 m^2g^{-1} by the six regression approaches. OLS yields the lowest slope (13.66 as shown in Figure 7a) among all six regression approaches, consistent with the results using synthetic data. This implies that OLS tends to underestimate regression slope when mean Y to X ratio is larger than 1.

DR ($\lambda = 1$) and ODR report the same slope (14.88) and intercept (5.54), this equivalency is also observed for the synthetic data. Similarly, WODR and YR yield identical slope (14.88) and intercept (5.54), in line with the synthetic data results. The regressed slope by DR ($\lambda = 1$) is higher than DR ($\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$), and this relationship agrees well with the synthetic data results.

Regression comparison is also performed on hourly OC and EC data. Regression on OC/EC percentile subset is a widely used empirical approach for primary OC/EC ratio determination. Figure S4 shows the regression slopes as a function of OC/EC percentile. OC/EC percentile ranges from 0.5% to 100%, with an interval of 0.5%. As the percentile increases, SOC contribution in OC increases as well, resulting decreased R^2 between OC and EC. The deviations between six regression approaches exhibit a dependency on R^2. When percentile is relatively small (e.g., <10%), the differences between the six regression approaches are also small due to the high R^2 (0.98). The deviations between the six regression approaches become more pronounced as R^2 decreases (e.g., <0.9). The deviations are expected to be even larger when R^2 is less than 0.8. These results emphasize the importance of applying error-in-variables regression, since ambient XY data more likely has a R^2 less than 0.9 in most cases.

As discussed in this section, the ambient data confirm the results obtained in comparing methods with the synthetic data. The advantage of using the synthetic data for regression approaches evaluation is that the ideal slope and intercept are known values during the data generation, so the bias of each regression approach can be quantified.

6 Recommendations and conclusions

This study aims to provide a benchmark of commonly used linear regression algorithms using a new data generation scheme (MT). Six regression approaches are tested, including OLS, DR ($\lambda = 1$), DR ($\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$), ODR, WODR and YR. The results show that OLS fails to estimate the correct slope and intercept when both X and Y have measurement errors. This result is consistent with previous studies. For ambient data with R^2 less than 0.9, error-in-variables regression is needed to minimize the biases in slope and intercept. If measurement uncertainties in X and Y are determined during the measurement, measurement uncertainties should be used for regression. With appropriate weighting, DR, WODR and YR can provide the best results among all tested regression techniques. Sensitivity tests also reveal the importance of the weighting parameter λ in DR. An improper λ could lead to biased slope and intercept. Since the λ estimation depends on the form of the measurement errors, it is important to determine the measurement errors during the experimentation stage rather than making assumptions. If measurement errors are not available from the measurement and assumptions are made on measurement errors, DR, WODR and YR are still the best option that can provide the least bias in slope and intercept among all tested regression techniques. For these reasons, DR, WODR and YR are recommended for atmospheric studies when both X and Y data have measurement errors.

Application of error-in-variables regression is often overlooked in atmospheric studies, partly due to the lack of a specified tool for the regression implementation. To facilitate the implementation of error-in-variables regression (including DR,WODR and YR), a computer program (Scatter plot) with graphical user interface (GUI) in Igor Pro (WaveMetrics, Inc. Lake Oswego, OR, USA) is developed (Figure 8). It packed with many useful features for data analysis and plotting, including batch plotting, data masking via GUI, color coding in Z axis, data filtering and grouping by numerical values and strings. The Scatter plot program and user manual are available from https://sites.google.com/site/wuchengust and https://doi.org/10.5281/zenodo.832417.

Appendix A: Equations of regression techniques

Ordinary Least Square (**OLS**) calculation steps.

First calculate average of observed X_i and Y_i.

$$\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N} \tag{A1}$$

$$\bar{Y} = \frac{\sum_{i=1}^{N} Y_i}{N} \tag{A2}$$

Then calculate S_{xx} and S_{yy}.

$$S_{xx} = \sum_{i=1}^{N} (X_i - \bar{X})^2 \tag{A3}$$

$$S_{yy} = \sum_{i=1}^{N} (Y_i - \bar{Y})^2 \tag{A4}$$

OLS slope and intercept can be obtained from,

$$k = \frac{S_{yy}}{S_{xx}} \tag{A6}$$

$$b = \bar{Y} - k\bar{X} \tag{A7}$$

Deming regression (**DR**) calculation steps (York, 1966).

Besides S_{xx} and S_{yy} as shown above, S_{xy} can be calculated from,

$$S_{xy} = \sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y}) \tag{A8}$$

DR slope and intercept can be obtained from,

$$k = \frac{S_{yy} - \lambda S_{xx} + \sqrt{\left(S_{yy} - \lambda S_{xx}\right)^2 + 4\lambda {S_{xy}}^2}}{2S_{xy}} \tag{A9}$$

$$b = \bar{Y} - k\bar{X} \tag{A10}$$

York regression (**YR**) iteration steps (York et al., 2004).

Slope by OLS can be used as the initial k in W_i calculation.

$$W_i = \frac{\omega(X_i)\omega(Y_i)}{\omega(X_i) + k^2\omega(Y_i) - 2kr_i\sqrt{\omega(X_i)\omega(Y_i)}} \tag{A11}$$

$$U_i = X_i - \bar{X} = X_i - \frac{\sum_{i=1}^{N} W_i X_i}{\sum_{i=1}^{N} W_i} \quad \text{(A12)}$$

$$V_i = Y_i - \bar{Y} = Y_i - \frac{\sum_{i=1}^{N} W_i Y_i}{\sum_{i=1}^{N} W_i} \quad \text{(A13)}$$

Then calculate β_i.

$$\beta_i = W_i \left[\frac{U_i}{\omega(Y_i)} + \frac{kV_i}{\omega(X_i)} - [kU_i + V_i] \frac{r_i}{\sqrt{\omega(X_i)\omega(Y_i)}} \right] \quad \text{(A14)}$$

Slope and intercept can be obtained from,

$$k = \frac{\sum_{i=1}^{n} W_i \beta_i V_i}{\sum_{i=1}^{n} W_i \beta_i U_i} \quad \text{(A15)}$$

$$b = \bar{Y} - k\bar{X} \quad \text{(A16)}$$

Since W_i and β_i are functions of k, k must be solved iteratively by repeating A11 to A15. If the difference between the k obtained from A15 and the k used in A11 satisfies the predefined tolerance ($\frac{k_{i+1} - k_i}{k_i} < e^{-15}$), the calculation is considered as converged. The calculation is straightforward and usually converged in 10 iterations. For example, the iteration count on the data set of Chu (2005) is around 6.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grant No. 41605002 and 21607056), NSFC of Guangdong Province (Grant No. 2015A030313339), Guangdong Province Public Interest Research and Capacity Building Special Fund (Grant No. 2014B020216005). The author would like to thank Dr. Bin Yu Kuang at HKUST for discussion on mathematics and Dr. Stephen M Griffith at HKUST for valuable comments.

References

Ayers, G. P.: Comment on regression analysis of air quality data, Atmos. Environ., 35, 2423-2425, doi: 10.1016/S1352-2310(00)00527-6, 2001.

Bauer, J. J., Yu, X.-Y., Cary, R., Laulainen, N., and Berkowitz, C.: Characterization of the sunset semi-continuous carbon aerosol analyzer, J. Air Waste Manage. Assoc., 59, 826-833, doi: 10.3155/1047-3289.59.7.826, 2009.

Boggs, P. T., Donaldson, J. R., and Schnabel, R. B.: Algorithm 676: ODRPACK: software for weighted orthogonal distance regression, ACM Trans. Math. Softw., 15, 348-364, doi: 10.1145/76909.76913, 1989.

Brauers, T. and Finlayson-Pitts, B. J.: Analysis of relative rate measurements, Int. J. Chem. Kinet., 29, 665-672, doi: 10.1002/(SICI)1097-4601(1997)29:9<665::AID-KIN3>3.0.CO;2-S, 1997.

Cantrell, C. A.: Technical Note: Review of methods for linear least-squares fitting of data and application to atmospheric chemistry problems, Atmos. Chem. Phys., 8, 5477-5487, doi: 10.5194/acp-8-5477-2008, 2008.

Carroll, R. J. and Ruppert, D.: The use and misuse of orthogonal regression in linear errors-in-variables models, Am. Stat., 50, 1-6, doi: 10.1080/00031305.1996.10473533, 1996.

Cess, R. D., Zhang, M. H., Minnis, P., Corsetti, L., Dutton, E. G., Forgan, B. W., Garber, D. P., Gates, W. L., Hack, J. J., Harrison, E. F., Jing, X., Kiehi, J. T., Long, C. N., Morcrette, J.-J., Potter, G. L., Ramanathan, V., Subasilar, B., Whitlock, C. H., Young, D. F., and Zhou, Y.: Absorption of solar radiation by clouds: Observations versus models, Science, 267, 496-499, doi: 10.1126/science.267.5197.496, 1995.

Chen, L. W. A., Doddridge, B. G., Dickerson, R. R., Chow, J. C., Mueller, P. K., Quinn, J., and Butler, W. A.: Seasonal variations in elemental carbon aerosol, carbon monoxide and sulfur dioxide: Implications for sources, Geophys. Res. Lett., 28, 1711-1714, doi: 10.1029/2000GL012354, 2001.

Chu, S. H.: Stable estimate of primary OC/EC ratios in the EC tracer method, Atmos. Environ., 39, 1383-1392, doi: 10.1016/j.atmosenv.2004.11.038, 2005.

Coen, M. C., Weingartner, E., Apituley, A., Ceburnis, D., Fierz-Schmidhauser, R., Flentje, H., Henzing, J. S., Jennings, S. G., Moerman, M., Petzold, A., Schmid, O., and Baltensperger, U.: Minimizing light absorption measurement artifacts of the Aethalometer: evaluation of five correction algorithms, Atmos. Meas. Tech., 3, 457-474, doi: 10.5194/amt-3-457-2010, 2010.

Cornbleet, P. J. and Gochman, N.: Incorrect least-squares regression coefficients in method-comparison analysis, Clin. Chem., 25, 432-438, 1979.

Cox, M., Harris, P., and Siebert, B. R.-L.: Evaluation of Measurement Uncertainty Based on the Propagation of Distributions Using Monte Carlo Simulation, Measurement Techniques, 46, 824-833, doi: 10.1023/B:METE.0000008439.82231.ad, 2003.

Cross, E. S., Onasch, T. B., Ahern, A., Wrobel, W., Slowik, J. G., Olfert, J., Lack, D. A., Massoli, P., Cappa, C. D., Schwarz, J. P., Spackman, J. R., Fahey, D. W., Sedlacek, A., Trimborn, A., Jayne, J. T., Freedman, A., Williams, L. R., Ng, N. L., Mazzoleni, C., Dubey, M., Brem, B., Kok, G., Subramanian, R., Freitag, S., Clarke, A., Thornhill, D., Marr, L. C., Kolb, C. E., Worsnop, D. R., and Davidovits, P.: Soot particle studies—instrument inter-comparison—project overview, Aerosol. Sci. Technol., 44, 592-611, doi: 10.1080/02786826.2010.482113, 2010.

Deming, W. E.: Statistical Adjustment of Data, Wiley, New York, 1943.

Duan, F., Liu, X., Yu, T., and Cachier, H.: Identification and estimate of biomass burning contribution to the urban aerosol organic carbon concentrations in Beijing, Atmos. Environ., 38, 1275-1282, doi: 10.1016/j.atmosenv.2003.11.037, 2004.

Flanagan, J. B., Jayanty, R. K. M., Rickman, J. E. E., and Peterson, M. R.: PM2.5 Speciation Trends Network: Evaluation of Whole-System Uncertainties Using Data from Sites with Collocated Samplers, J. Air Waste Manage. Assoc., 56, 492-499, doi: 10.1080/10473289.2006.10464516, 2006.

Hansen, A. D. A.: The Aethalometer Manual, Berkeley, California,USA, Magee Scientific, 2005.

Huang, X. H., Bian, Q., Ng, W. M., Louie, P. K., and Yu, J. Z.: Characterization of PM2. 5 major components and source investigation in suburban Hong Kong: A one year monitoring study, Aerosol. Air. Qual. Res., 14, 237-250, doi: 10.4209/aaqr.2013.01.0020, 2014.

Janhäll, S., Andreae, M. O., and Pöschl, U.: Biomass burning aerosol emissions from vegetation fires: particle number and mass emission factors and size distributions, Atmos. Chem. Phys., 10, 1427-1439, doi: 10.5194/acp-10-1427-2010, 2010.

Lim, L. H., Harrison, R. M., and Harrad, S.: The contribution of traffic to atmospheric concentrations of polycyclic aromatic hydrocarbons, Environ. Sci. Technol., 33, 3538-3542, doi: 10.1021/es990392d, 1999.

Linnet, K.: Necessary sample size for method comparison studies based on regression analysis, Clin. Chem., 45, 882-894, 1999.

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.

Markovsky, I. and Van Huffel, S.: Overview of total least-squares methods, Signal Process., 87, 2283-2302, doi: 10.1016/j.sigpro.2007.04.004, 2007.

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

Moosmüller, H., Arnott, W. P., Rogers, C. F., Chow, J. C., Frazier, C. A., Sherman, L. E., and Dietrich, D. L.: Photoacoustic and filter measurements related to aerosol light absorption during the Northern Front Range Air Quality Study (Colorado 1996/1997), J. Geophys. Res., 103, 28149-28157, doi: 10.1029/98jd02618, 1998.

Petäjä, T., Mauldin, I. R. L., Kosciuch, E., McGrath, J., Nieminen, T., Paasonen, P., Boy, M., Adamov, A., Kotiaho, T., and Kulmala, M.: Sulfuric acid and OH concentrations in a boreal forest site, Atmos. Chem. Phys., 9, 7435-7448, doi: 10.5194/acp-9-7435-2009, 2009.

Richter, A., Burrows, J. P., Nusz, H., Granier, C., and Niemeier, U.: Increase in tropospheric nitrogen dioxide over China observed from space, Nature, 437, 129-132, doi: 10.1038/nature04092, 2005.

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.

Thompson, M.: Variation of precision with concentration in an analytical system, Analyst, 113, 1579-1587, doi: 10.1039/AN9881301579, 1988.

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, doi: 10.1016/1352-2310(94)00276-Q, 1995.

von Bobrutzki, K., Braban, C. F., Famulari, D., Jones, S. K., Blackall, T., Smith, T. E. L., Blom, M., Coe, H., Gallagher, M., Ghalaieny, M., McGillen, M. R., Percival, C. J., Whitehead, J. D., Ellis, R., Murphy, J., Mohacsi, A., Pogany, A., Junninen, H., Rantanen, S., Sutton, M. A., and Nemitz, E.: Field inter-comparison of eleven atmospheric ammonia measurement techniques, Atmos. Meas. Tech., 3, 91-112, doi: 10.5194/amt-3-91-2010, 2010.

Wang, J. and Christopher, S. A.: Intercomparison between satellite-derived aerosol optical thickness and PM2.5 mass: Implications for air quality studies, Geophys. Res. Lett., 30, 2095, doi: 10.1029/2003gl018174, 2003.

Watson, J. G.: Visibility: Science and regulation, J. Air Waste Manage. Assoc., 52, 628-713, doi: 10.1080/10473289.2002.10470813, 2002.

Weingartner, E., Saathoff, H., Schnaiter, M., Streit, N., Bitnar, B., and Baltensperger, U.: Absorption of light by soot particles: determination of the absorption coefficient by means of aethalometers, J. Aerosol. Sci., 34, 1445-1463, doi: 10.1016/S0021-8502(03)00359-8, 2003.

Wu, C. and Yu, J. Z.: 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, Atmos. Chem. Phys., 16, 5453-5465, doi: 10.5194/acp-16-5453-2016, 2016.

Wu, C., Wu, D., and Yu, J. Z.: Quantifying black carbon light absorption enhancement by a novel statistical approach, Atmos. Chem. Phys. Discuss., 2017, 1-37, doi: 10.5194/acp-2017-582, 2017.

York, D.: Least-squares fitting of a straight line, Can. J. Phys., 44, 1079-1086, doi: 10.1139/p66-090, 1966.

York, D., Evensen, N. M., Martınez, M. L., and Delgado, J. D. B.: Unified equations for the slope, intercept, and standard errors of the best straight line, Am. J. Phys., 72, 367-375, doi: 10.1119/1.1632486, 2004.

Zhou, Y., Huang, X. H. H., Griffith, S. M., Li, M., Li, L., Zhou, Z., Wu, C., Meng, J., Chan, C. K., Louie, P. K. K., and Yu, J. Z.: A field measurement based scaling approach for quantification of major ions, organic carbon, and elemental carbon using a single particle aerosol mass spectrometer, Atmos. Environ., 143, 300-312, doi: 10.1016/j.atmosenv.2016.08.054, 2016.

Zieger, P., Weingartner, E., Henzing, J., Moerman, M., de Leeuw, G., Mikkilä, J., Ehn, M., Petäjä, T., Clémer, K., van Roozendael, M., Yilmaz, S., Frieß, U., Irie, H., Wagner, T., Shaiganfar, R., Beirle, S., Apituley, A., Wilson, K., and Baltensperger, U.: Comparison of ambient aerosol extinction coefficients obtained from in-situ, MAX-DOAS and LIDAR measurements at Cabauw, Atmos. Chem. Phys., 11, 2603-2624, doi: 10.5194/acp-11-2603-2011, 2011.

Zwolak, J. W., Boggs, P. T., and Watson, L. T.: Algorithm 869: ODRPACK95: A weighted orthogonal distance regression code with bound constraints, ACM Trans. Math. Softw., 33, 27, doi: 10.1145/1268776.1268782, 2007.

Table 1. Summary of abbreviations and symbols.

Abbreviation/symbol	Definition
α	a dimensionless adjustable factor to control the position of γ_{Unc} curve on the concentration axis
b	intercept in linear regression
β_i, U_i, V_i, W_i	intermediates in York regression calculations
γ_{Unc}	fractional measurement uncertainties relative to the true concentration (%)
DR	Deming regression
ε_{EC} , ε_{POC}	absolute measurement uncertainties of EC and POC
EC	elemental carbon
EC_{true}	numerically synthesized true EC concentration without measurement uncertainty
$EC_{measured}$	EC with measurement error ($EC_{true} + \varepsilon_{EC}$)
λ	$\omega(X_i)$ to $\omega(Y_i)$ ratio in Deming regression
k	slope in linear regression
LOD	limit of detection
MT	Mersenne twister pseudorandom number generator
OC	organic carbon
OC/EC	OC to EC ratio
$(OC/EC)_{pri}$	primary OC/EC ratio
$OC_{non\text{-}comb}$	OC from non-combustion sources
ODR	orthogonal distance regression
OLS	ordinary least squares regression
POC	primary organic carbon
POC_{comb}	numerically synthesized true POC from combustion sources (well correlated with EC_{true}), measurement uncertainty not considered
$POC_{non\text{-}comb}$	numerically synthesized true POC from non-combustion sources (independent of EC_{true}) without considering measurement uncertainty
POC_{true}	sum of POC_{comb} and $POC_{non\text{-}comb}$ without considering measurement uncertainty
$POC_{measured}$	POC with measurement error ($POC_{true} + \varepsilon_{POC}$)
σ_{X_i} , σ_{Y_i}	the standard deviation of the error in measurement of X_i and Y_i
r_i	correlation coefficient between errors in X_i and Y_i in YR
S	sum of squared residuals
SOC	secondary organic carbon
τ	parameter in the sine function of Chu (2005) that adjust the width of each peak
ϕ	parameter in the sine function of Chu (2005) that adjust the phase of the curve
WODR	weight orthogonal distance regression
$\bar{X}$, $\bar{Y}$	average of X_i and Y_i
YR	York regression
$\omega(X_i)$, $\omega(Y_i)$	inverse of σ_{X_i} and σ_{Y_i}, used as weights in DR calculation.

Table 2. Summary of six regression approaches comparison with 5000 runs for 18 cases.

Data generation						Results by different regression approaches											
						OLS		DR λ=1		DR $\lambda=\frac{\omega(X_i)}{\omega(Y_i)}$		ODR		WODR		YR	
Case	Data scheme	True Slope	True Intercept	R^2 (X, Y)	Measurement error	Slope	Intercept	Slope	Intercept	Slope	Intercept	Slope	Intercept	Slope	Intercept	Slope	Intercept
1	Chu	4	0	0.67±0.03	LOD_{POC}=1, LOD_{EC}=1 α_{POC}=1, α_{EC}=1.	2.94±0.14	5.84±0.78	4.27±0.27	-1.45±1.36	4.01±0.25	-0.04±1.28	4.27±0.27	-1.45±1.36	3.98±0.22	1.12±1.02	3.98±0.22	1.12±1.02
2		4	3	0.67±0.04		2.95±0.15	8.83±0.80	4.32±0.28	1.28±1.43	4.01±0.26	2.94±1.34	4.32±0.28	1.28±1.43	3.99±0.23	3.98±1.05	3.99±0.23	3.98±1.05
3		4	0	0.95±0.01	LOD_{POC}=0.5, LOD_{EC}=0.5 α_{POC}=0.5, α_{EC}=0.5	3.83±0.08	0.95±0.40	4.03±0.09	-0.18±0.44	4±0.09	0±0.44	4.03±0.09	-0.18±0.44	4±0.08	0.12±0.37	4±0.08	0.12±0.37
4		4	0	0.78±0.02	LOD_{POC}=1, LOD_{EC}=0.5 α_{POC}=1, α_{EC}=1	3.39±0.15	3.34±0.75	4.3±0.21	-1.66±1.06	4±0.19	-0.03±0.99	4.3±0.21	-1.66±1.06	4±0.17	0.33±0.81	4±0.17	0.33±0.81
5		4	0	0.69±0.04	γ_{Unc}=30%	3.32±0.20	3.77±0.90	4.75±0.30	-4.14±1.36	4.01±0.25	-0.04±1.13	4.75±0.30	-4.14±1.36	4±0.18	-0.01±0.59	4±0.18	-0.01±0.59
6		4	3	0.66±0.04		3.31±0.22	6.79±1.02	4.95±0.31	-2.26±1.48	3.99±0.26	3.05±1.22	4.95±0.31	-2.26±1.48	4.01±0.20	2.72±0.74	4.01±0.20	2.72±0.74
7	MT	4	0	0.76±0.01	LOD_{POC}=1, LOD_{EC}=1 α_{POC}=1, α_{EC}=1	3.22±0.03	4.3±0.14	4.17±0.04	-0.94±0.18	4±0.03	0±0.17	4.17±0.04	-0.94±0.18	3.96±0.03	1.21±0.13	3.96±0.03	1.21±0.13
8		4	3	0.75±0.01		3.22±0.03	7.29±0.14	4.2±0.04	1.88±0.18	4±0.03	3±0.18	4.2±0.04	1.88±0.18	3.97±0.03	4.11±0.13	3.97±0.03	4.11±0.13
9		0.5	0	0.76±0.01		0.43±0.00	0.36±0.02	0.46±0.01	0.23±0.03	0.5±0.01	0±0.03	0.46±0.01	0.23±0.03	0.5±0.00	0±0.01	0.5±0.00	0±0.01
10		0.5	3	0.56±0.01		0.43±0.01	3.36±0.03	0.5±0.01	3.02±0.04	0.49±0.01	3.05±0.04	0.5±0.01	3.02±0.04	0.51±0.01	2.73±0.03	0.51±0.01	2.73±0.03
11		1	0	0.76±0.01		0.87±0.01	0.72±0.05	1±0.01	0±0.06	1±0.01	0±0.06	1±0.01	0±0.06	1±0.01	0±0.02	1±0.01	0±0.02
12		1	3	0.66±0.01		0.87±0.01	3.72±0.05	1.09±0.01	2.52±0.07	0.99±0.01	3.07±0.06	1.09±0.01	2.52±0.07	1.01±0.01	2.71±0.04	1.01±0.01	2.7±0.04
13		4	0	0.76±0.01	γ_{Unc}=30%	3.48±0.04	2.87±0.18	4.53±0.05	-2.94±0.24	4±0.05	0±0.22	4.53±0.05	-2.94±0.24	4±0.03	0±0.09	4±0.03	0±0.09
14		4	3	0.73±0.01		3.48±0.04	5.87±0.19	4.67±0.05	-0.67±0.26	3.98±0.05	3.08±0.23	4.67±0.05	-0.67±0.26	4.02±0.03	2.68±0.11	4.02±0.03	2.68±0.11
15		0.5	0	0.54±0.01		0.4±0.01	0.55±0.03	0.45±0.01	0.26±0.03	0.5±0.01	0.01±0.03	0.45±0.01	0.26±0.03	0.52±0.01	-0.23±0.02	0.52±0.01	-0.23±0.02
16		0.5	3	0.40±0.01		0.4±0.01	3.54±0.04	0.5±0.01	2.98±0.04	0.5±0.01	3±0.04	0.5±0.01	2.98±0.04	0.52±0.01	2.65±0.04	0.52±0.01	2.65±0.04
17		1	0	0.65±0.01		0.8±0.01	1.07±0.04	1±0.01	0±0.05	1±0.01	0±0.05	1±0.01	0±0.05	1±0.01	0±0.04	1±0.01	0±0.04
18		1	3	0.59±0.01		0.8±0.01	4.07±0.05	1.07±0.01	2.62±0.07	1±0.01	3±0.06	1.07±0.01	2.62±0.07	1.02±0.01	2.84±0.05	1.02±0.01	2.84±0.05

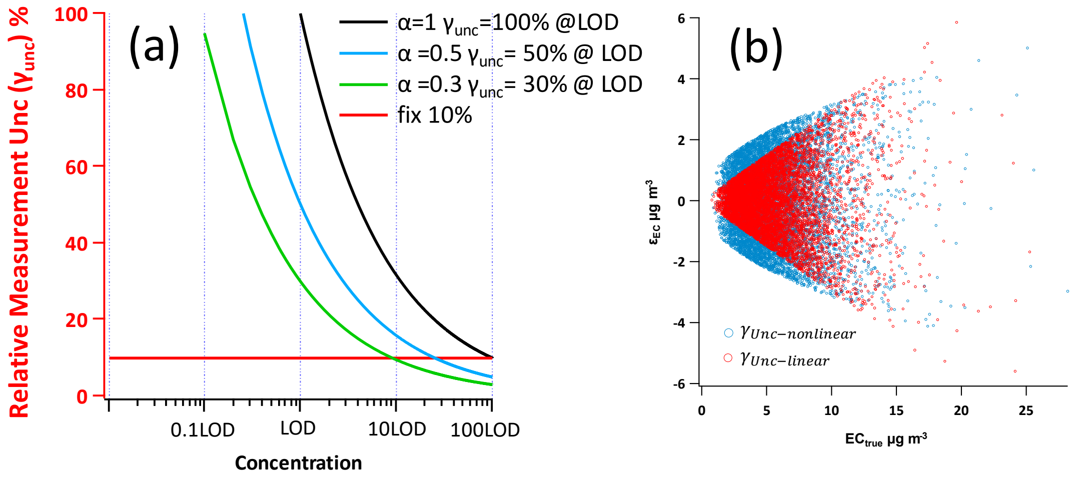
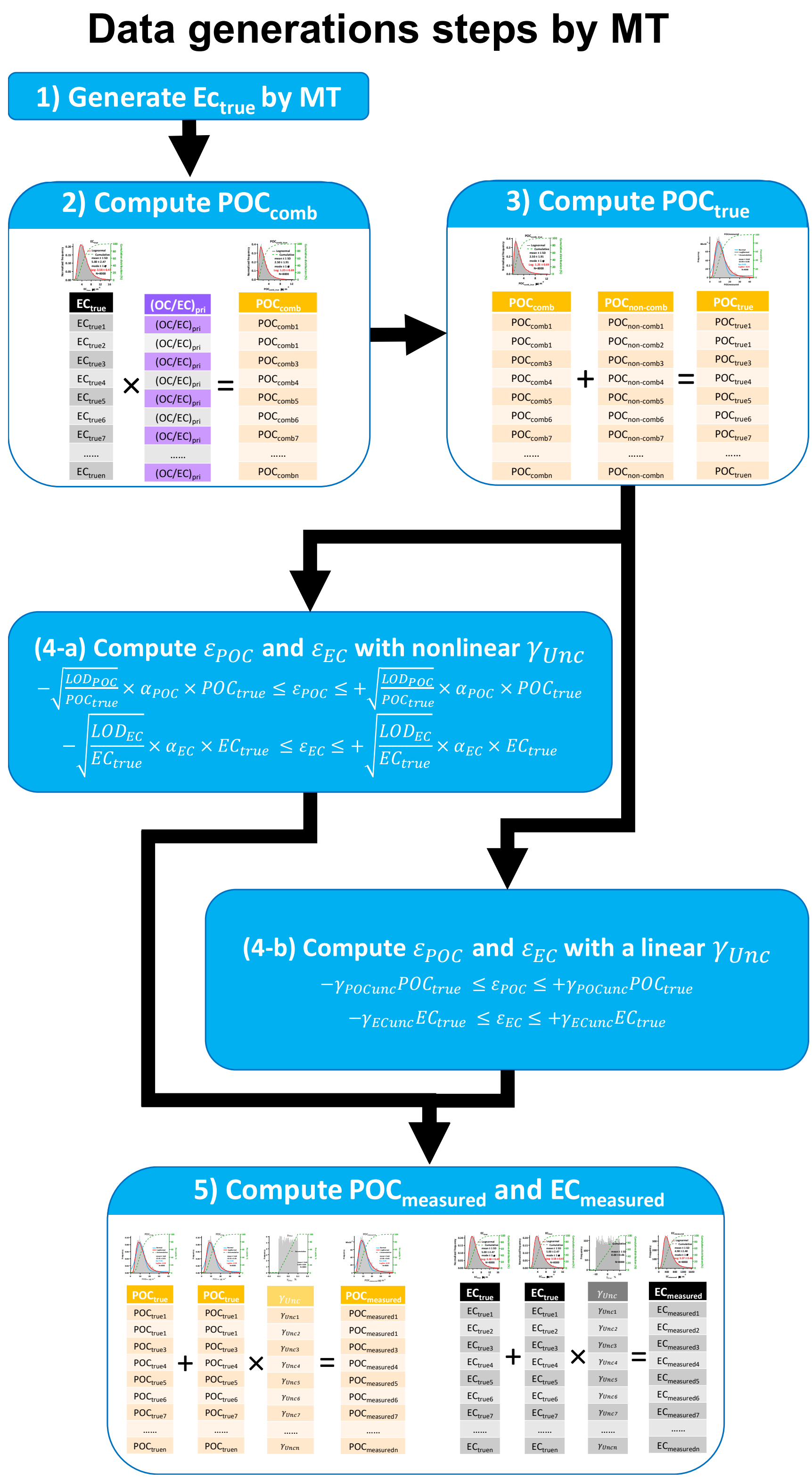


Figure 1. (a) Example $\gamma_{Unc-nonlinear}$ curves by different α values (Eq. (17)). The X axis is concentration (normalized by LOD) in log scale and Y axis is γ_{Unc}. Black, blue and green line represent α equal to 1, 0.5 and 0.3 respectively, corresponding to the $\gamma_{Unc-nonlinear}$ at LOD level equals to 100%, 50% and 30% respectively. The red line represents $\gamma_{Unc-linear}$ of 10%. (b) Example of measurement uncertainty generation of $\gamma_{Unc-nonlinear}$ and $\gamma_{Unc-linear}$. The blue circles represent $\gamma_{Unc-nonlinear}$ following Eq. (17) ($LOD_{EC} = 1$, $a_{EC} = 1$). The red circles represent $\gamma_{Unc-linear}$ (30%).

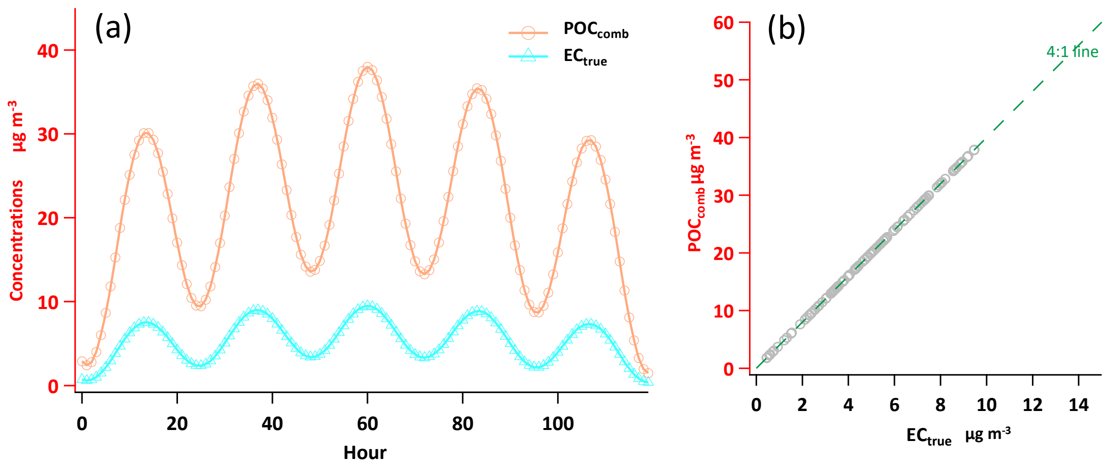
772



773

774 **Figure 2.** Flowchart of data generation steps using MT.

775



$$POC_{comb} = 14 + 12(\sin\left(\frac{x}{\tau}\right) + \sin(x - \phi))$$

$$EC_{true} = 3.5 + 3(\sin\left(\frac{x}{\tau}\right) + \sin(x - \phi))$$

Figure 3. POC_{comb} and EC_{trure} data generated by the sine functions of (Chu (2005)). (a) Time series of the 120 data points for POC_{comb} and EC_{true}. (b) Scatter plot of POC_{comb} vs. EC_{true}

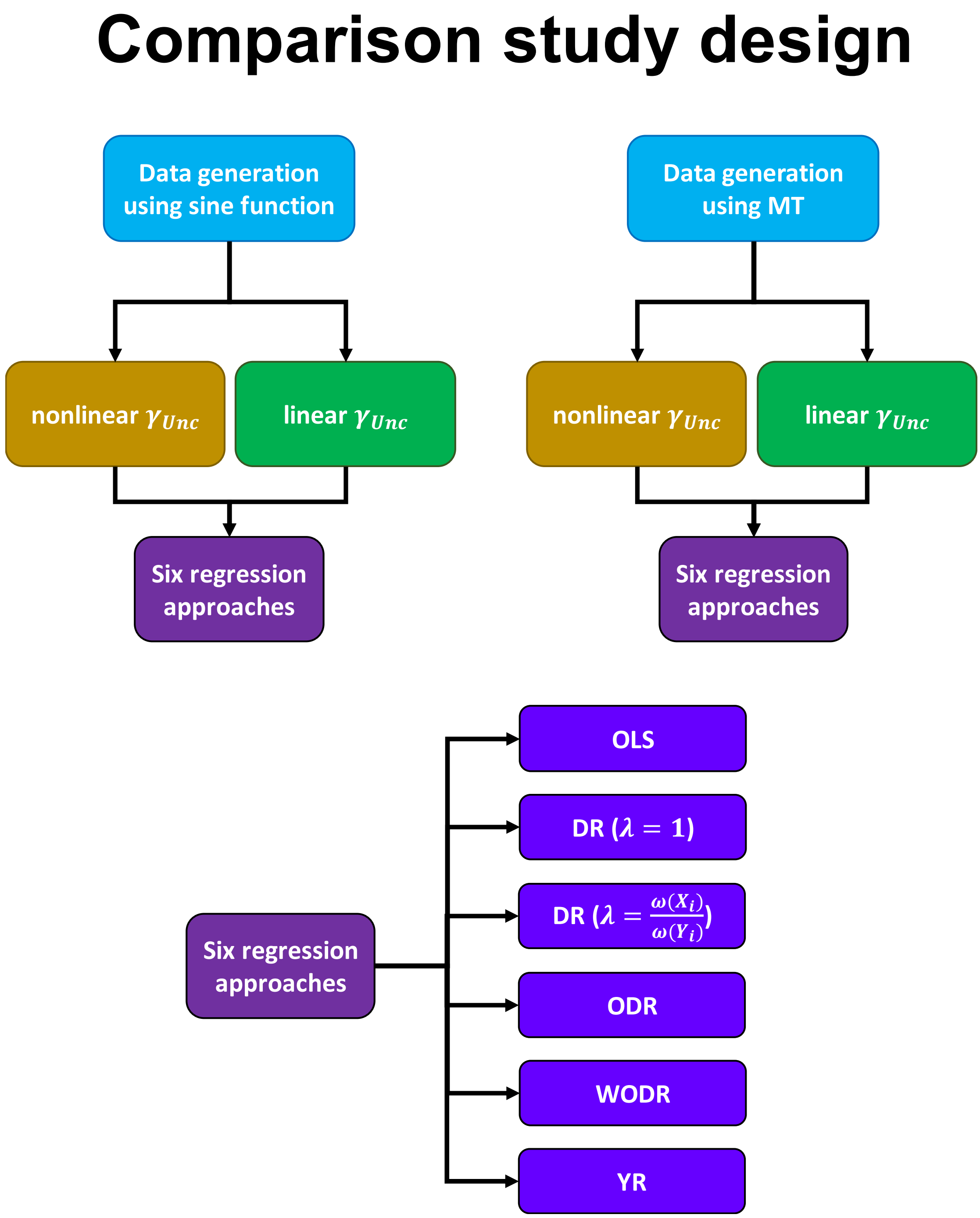


Figure 4. Overview of the comparison study design.

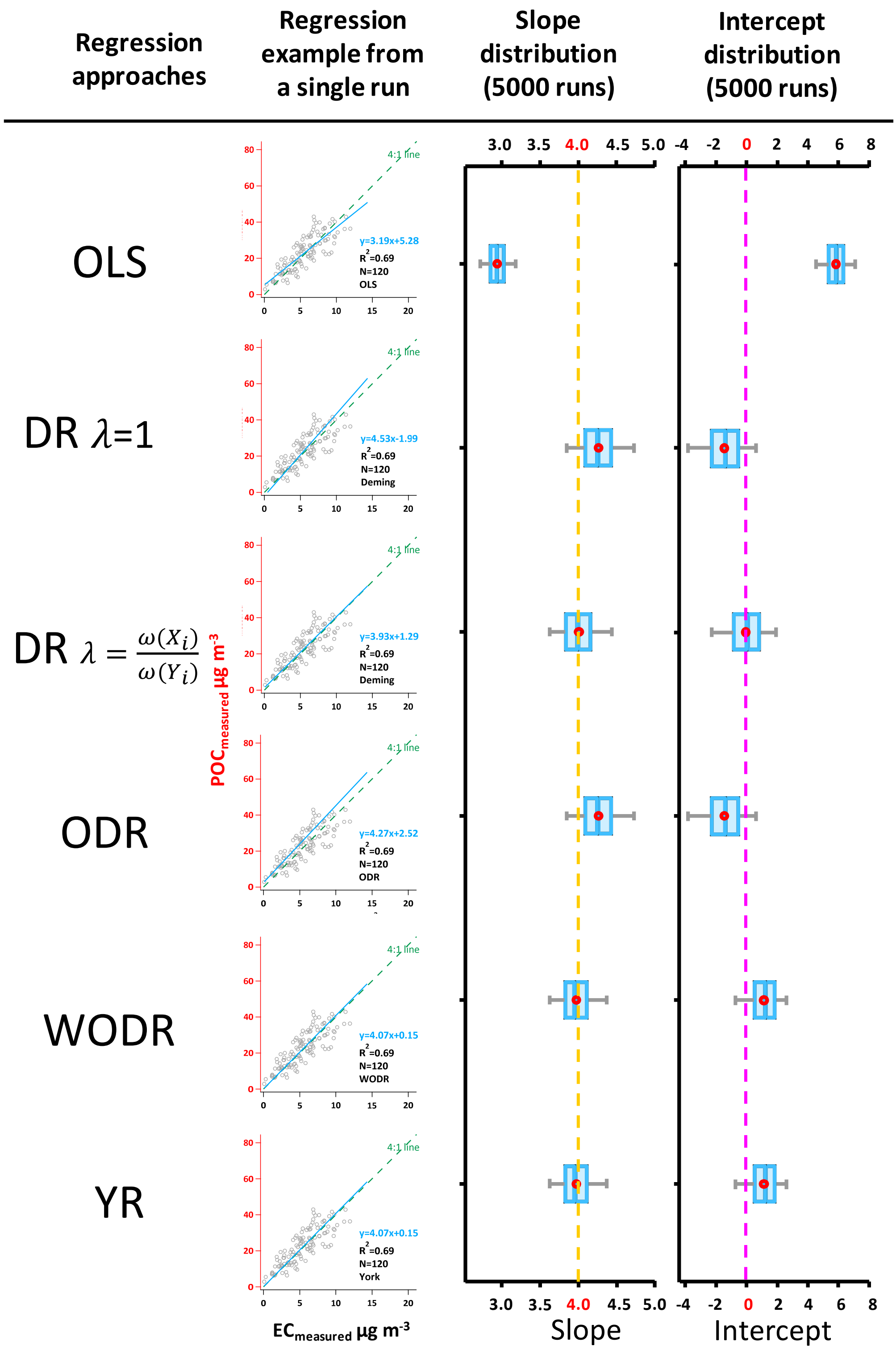


Figure 5. Regression results on synthetic data, case 1 (Slope=4, Intercept=0, LOD_{POC}=1, LOD_{EC}=1, a_{POC}=1, a_{EC}=1, R^2 (POC, EC) =0.67±0.03). The scatter plots demonstrate regression examples from a single run. The box plots show the distribution of regressed slopes and intercepts from 5000 runs of six regression approaches. The dashed line in orange and peachblow represent true slope and intercept respectively.

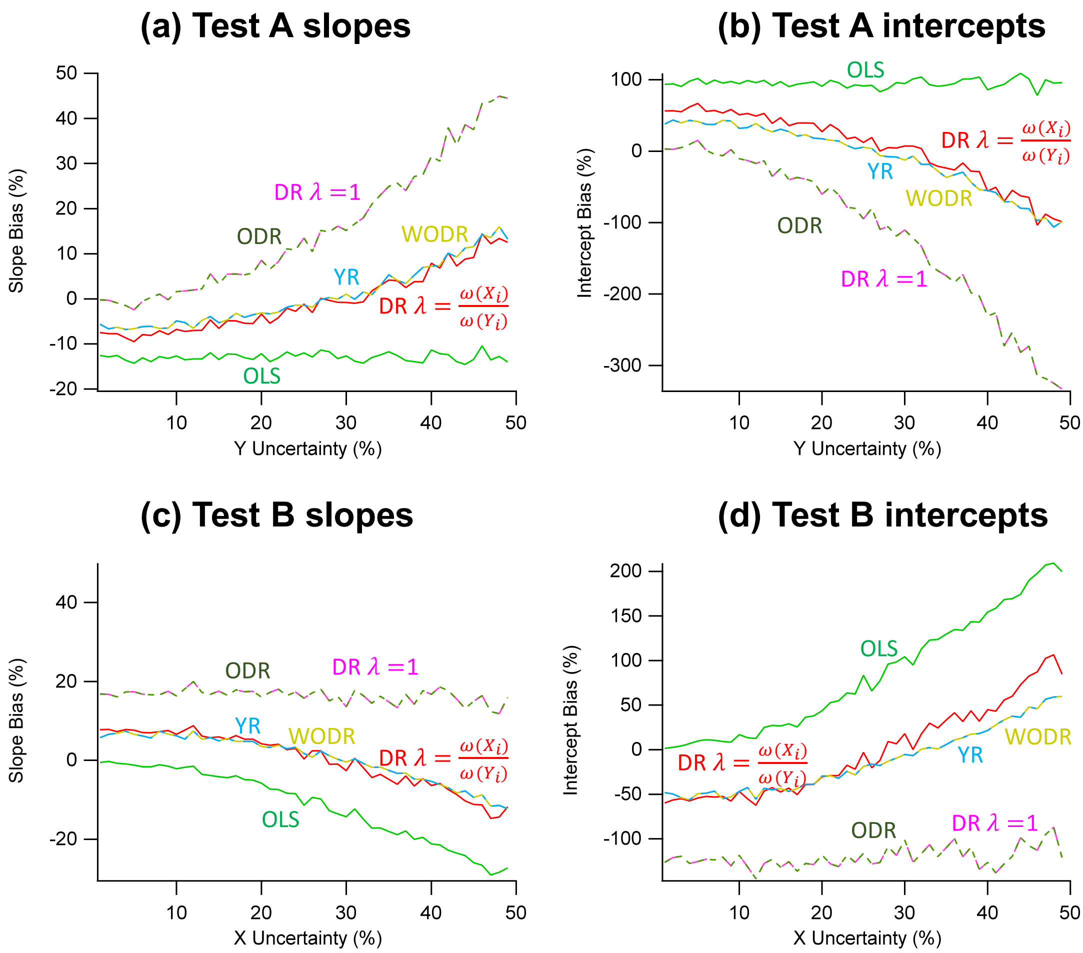


Figure 6. Slope and intercept biases due to the inconsistency between measurement error of data and measurement error used in regression. In Test A data generation, γ_{Unc_X} is fixed at 30% and γ_{Unc_Y} varied between 1 ~ 50%. In Test B, γ_{Unc_X} varied between 1 ~ 50% and γ_{Unc_Y} is fixed at 30%. The assumed measurement error for regression is 10% for both X and Y. (a) Slopes biases as a function of γ_{Unc_Y} in Test A. (b) Intercepts biases as a function of γ_{Unc_Y} in Test A. (c) Slopes biases as a function of γ_{Unc_X} in Test B. (d) Intercepts biases as a function of γ_{Unc_X} in Test B.

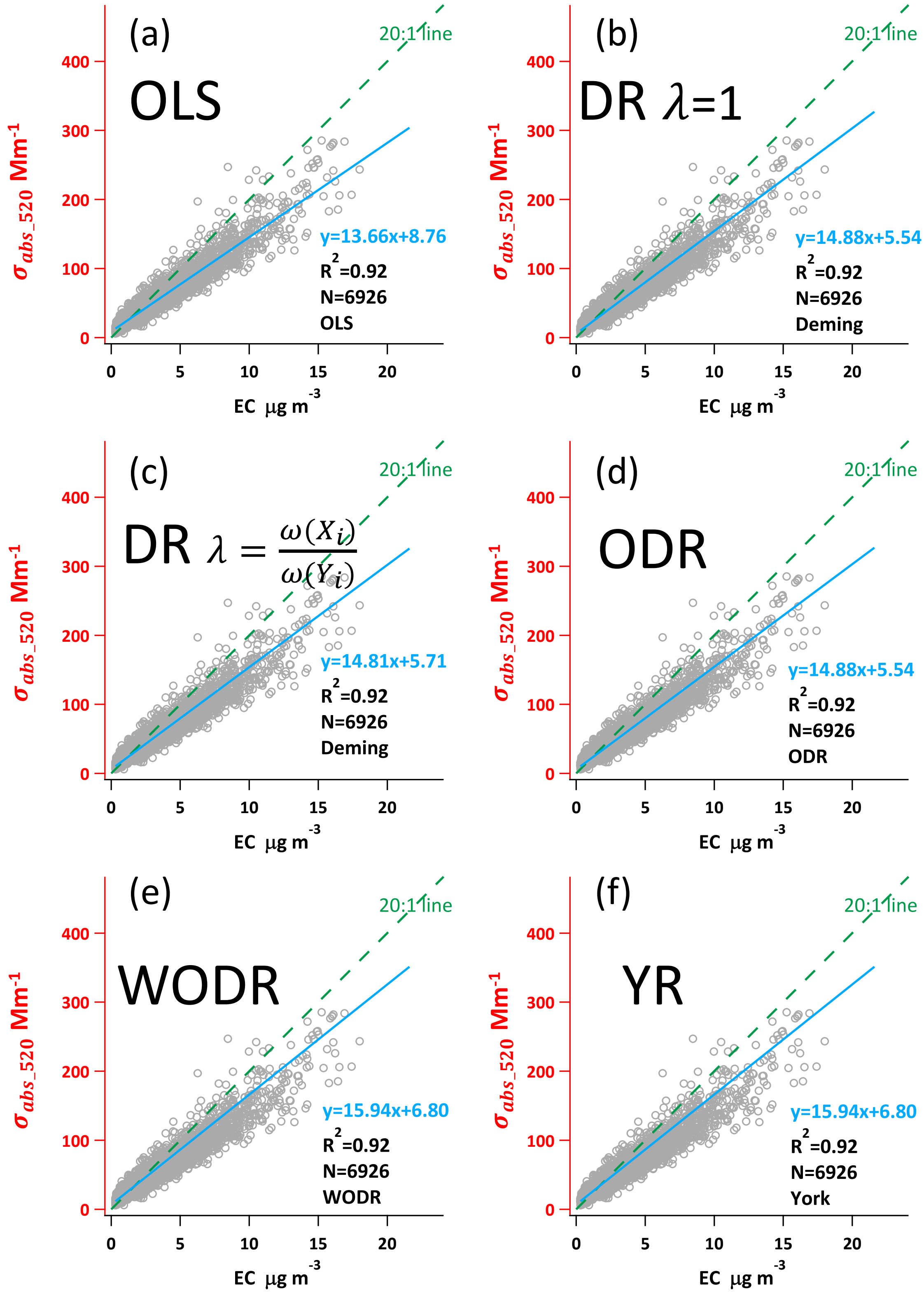


Figure 7. Regression results using ambient σ_{abs520} and EC data from a suburban site in Guangzhou, China.

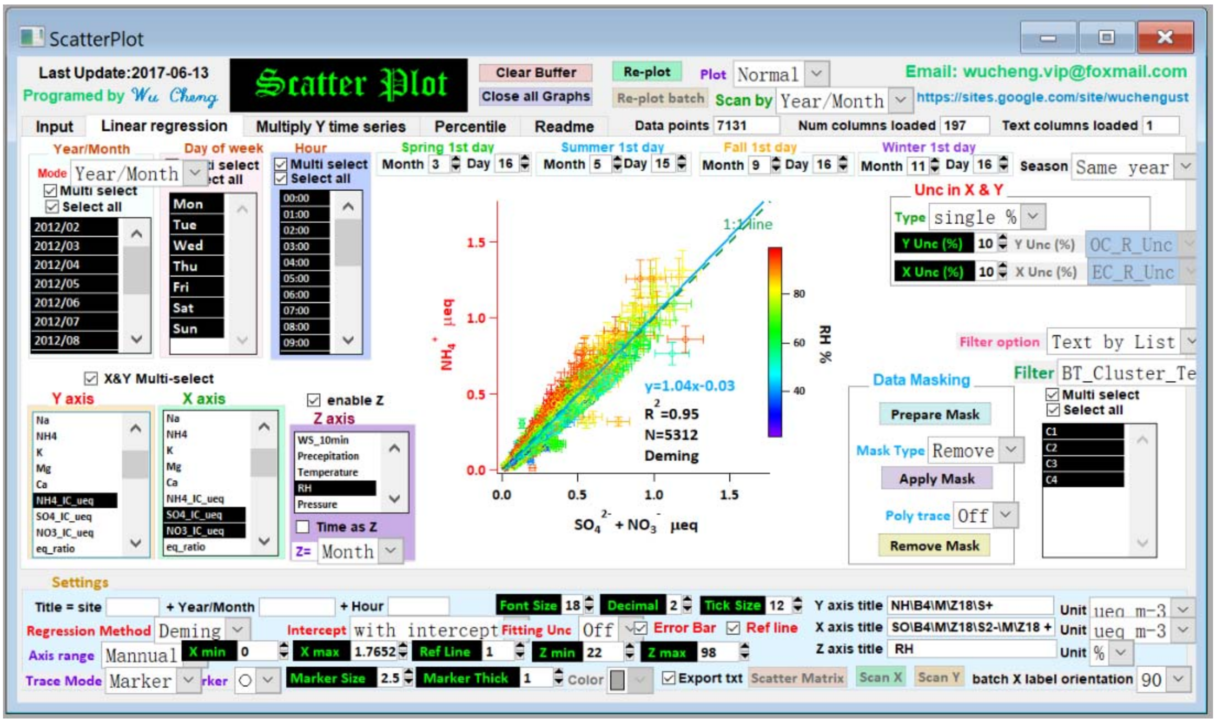


Figure 8. The user interface of Scatter Plot Igor program. The program and its operation manual are available from: https://doi.org/10.5281/zenodo.832417.

Supplement of

Evaluation of linear regression techniques for atmospheric applications: the importance of appropriate weighting

Cheng Wu[1,2] and Jian Zhen Yu[3,4,5]

[1]Institute of Mass Spectrometer and Atmospheric Environment, Jinan University, Guangzhou 510632, China

[2]Guangdong Provincial Engineering Research Center for on-line source apportionment system of air pollution, Guangzhou 510632, China

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

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

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

Corresponding to: Cheng Wu (wucheng.vip@foxmail.com) and Jian Zhen Yu (jian.yu@ust.hk)

This document contains two supporting tables, eight supporting figures.

1 Comparion of three York regression implementations

A variety of York regression implementations are compared using the Pearson's data with York's weights according to York (1966) (abbreviated as "PY data" hereafter). The dataset is Table S1.Three York regression implementations are compared using the PY data, including spreadsheet by Cantrell (2008), Igor program by this study and a commercial software (OriginPro™ 2017). The three York regression implementations yield identical slope and intercept as shown in the highlighted areas (in red) in Figure S5. These crosscheck results suggest that the codes in our Igor program can retrieve consistent slopes and intercepts as other proven programs did.

2 Impact of two primary sources in OC/EC regression

A sampling site is often impacted by multiple combustion sources in the real atmosphere. In section 1 and 2 we evaluate the performance of OLS, DR, WODR and YR 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. Two scenarios are considered (Wu and Yu, 2016): two correlated primary sources and two independent primary sources. Common configurations include: EC_{total}=2 μgC m^{-3}; f_{EC1} varies from 0 to 100%; ratio of the two OC/EC_{pri} values (γ_{pri}) vary in the range of 2~8. Studies by Chu (2005) and Saylor et al. (2006) both suggest ROA being the best estimator of the expected primary OC/EC ratio when SOC is zeroed. Since the overall OC/EC_{pri} from the two source varis by γ_{pri}, ROA is considered as the reference OC/EC_{pri} to be compared with slope regressed by of OLS, DR, WODR and YR. The abbreviations used for two primary sources study are listted in Table S8.

2.1 Impact of two correlated primary sources

Simulations considering two correlated primary sources are performed, to examine the effect on bias in the regression methods. The basic configuration is: $(OC/EC)_{pri1}$=0.5, $(OC/EC)_{pri2}$=5, γ_{Unc}=30%, N=8000, intercept=0, and the following terms are compared: ratio of average (ROA, which is considered as the true value of slope when intercept=0), DR, WODR, WODR' (through origin) and OLS. As shown in Figure S6, when R^2 (EC1 vs. EC2) is very high, DR, WODR and WODR' can provide a result consistent with ROA. If the R^2 decreases, the bias of the slope and intercept in DR and WODR is larger. OLS constantly underestimate the slope.

2.2 Impact of two independent primary sources

Simulations of two independent primary sources are also conducted. If $RSD_{EC1}=RSD_{EC2}$, slopes and intercepts may be either overestimated or underestimated (Figure S7), and the degree of bias depends on the magnitude of RSD_{EC1} and RSD_{EC2}. Larger RSD results in larger bias. Uneven RSD between two sources leads to even more bias (Figure S7 a&b). The degree of bias also shows dependence on γ_pri. If γ_pri decreases, the bias becomes smaller (FigureS7 c~f). These results indicate that the scenario with two independent primary sources poses a challenge to $(OC/EC)_{pri}$ estimation by linear regression.

For the EC tracer method, if EC comes from two primary sources and contribution of the two sources is comparable, the regression slope is no longer suitable for $(OC/EC)_{pri}$ estimation and the subsequent SOC calculation, and making EC a mixture that violates the property of a tracer. For such a situation, pre-separation of EC into individual sources by other tracers (if available) by the Minimum R Squared (MRS) method can provide unbiased SOC estimation results (Wu and Yu, 2016).

3 Igor programs for error in variables linear regression and simulated OC EC data generation using MT

An Igor Pro (WaveMetrics, Inc. Lake Oswego, OR, USA) based program (Scatter plot) with graphical user interface (GUI) is developed to make the linear regression feasible and user friendly (Figure 8). The program includes Deming and York algorithm for linear regression, which consider uncertainties in both X and Y, that is more realistic for atmospheric applications. It packed with many useful features for data analysis and plotting, including batch plotting, data masking via GUI, color coding in Z axis, data filtering and grouping by numerical values and strings.

Another program using MT can generate simulated OC and EC concentration through user defined parameters via GUI as shown in Figure S8.

Both Igor programs and their operation manuals can be downloaded from the following links:

https://sites.google.com/site/wuchengust

https://doi.org/10.5281/zenodo.832417

References

Chu, S. H.: Stable estimate of primary OC/EC ratios in the EC tracer method, Atmos. Environ., 39, 1383-1392, 10.1016/j.atmosenv.2004.11.038, 2005.

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, 10.1016/j.atmosenv.2006.07.018, 2006.

Wu, C. and Yu, J. Z.: 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, Atmos. Chem. Phys., 16, 5453-5465, 10.5194/acp-16-5453-2016, 2016.

Table S1. Pearson's data with York's weights according to York (1966).

X_i	$\omega(X_i)$	Y_i	$\omega(Y_i)$
0	1000	5.9	1
0.9	1000	5.4	1.8
1.8	500	4.4	4
2.6	800	4.6	8
3.3	200	3.5	20
4.4	80	3.7	20
5.2	60	2.8	70
6.1	20	2.8	70
6.5	1.8	2.4	100
7.4	1	1.5	500

Table S2. Abbreviations used for two primary sources study.

Abbreviation	**Definition**
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
ROA	ratio of averages
γ_pri	ratio of the $(OC/EC)_{pri}$ of source 2 to source 1
RSD	relative standard deviation
RSD_{EC}	RSD of EC
ε_{EC} , ε_{OC}	measurement uncertainty of EC and OC
γ_{unc}	relative measurement uncertainty
γ_{RSD}	the ratio between the RSD values of $(OC/EC)_{pri}$ and EC

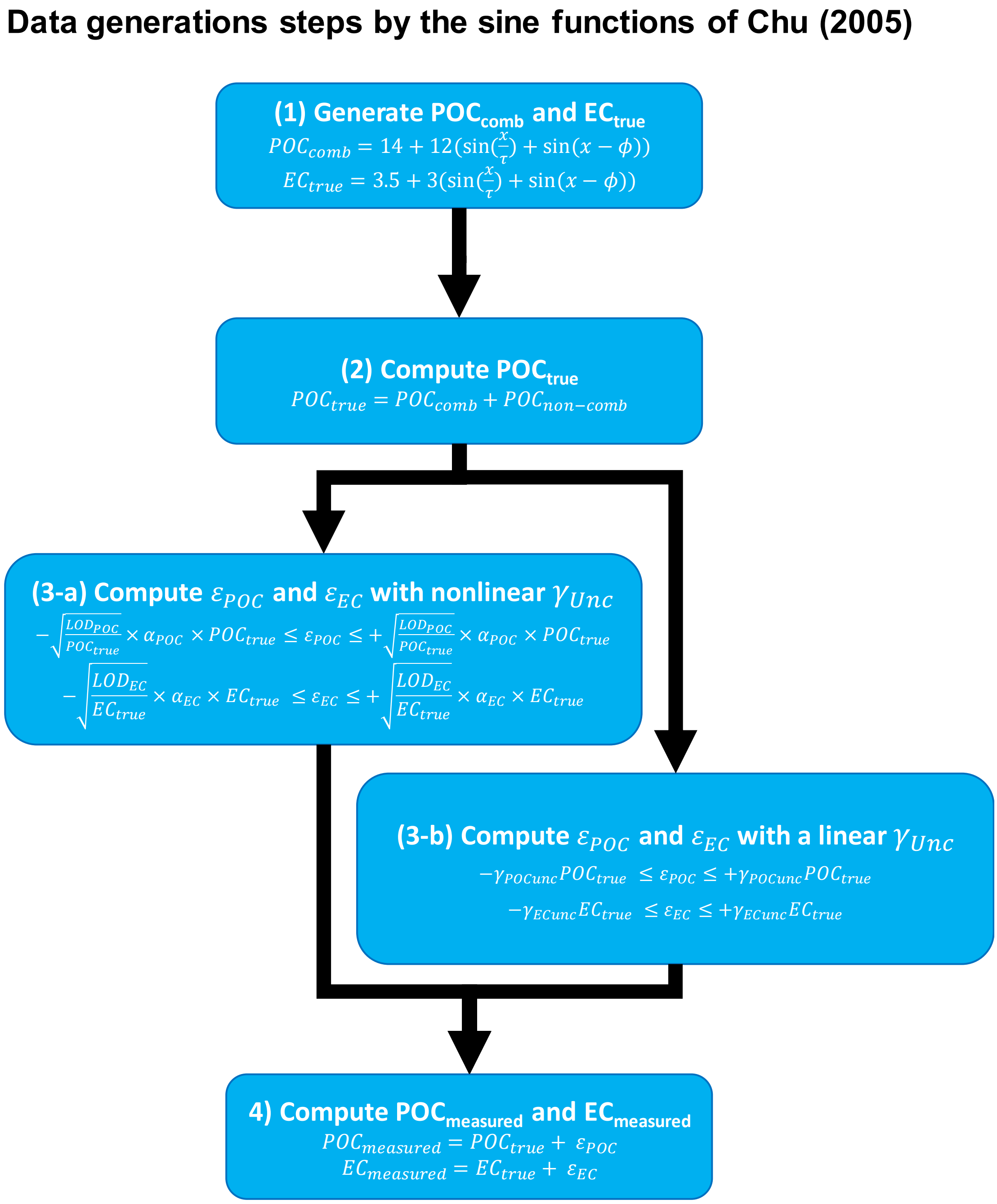


Figure S1. Flowchart of data generation steps using the sine functions of Chu (2005).

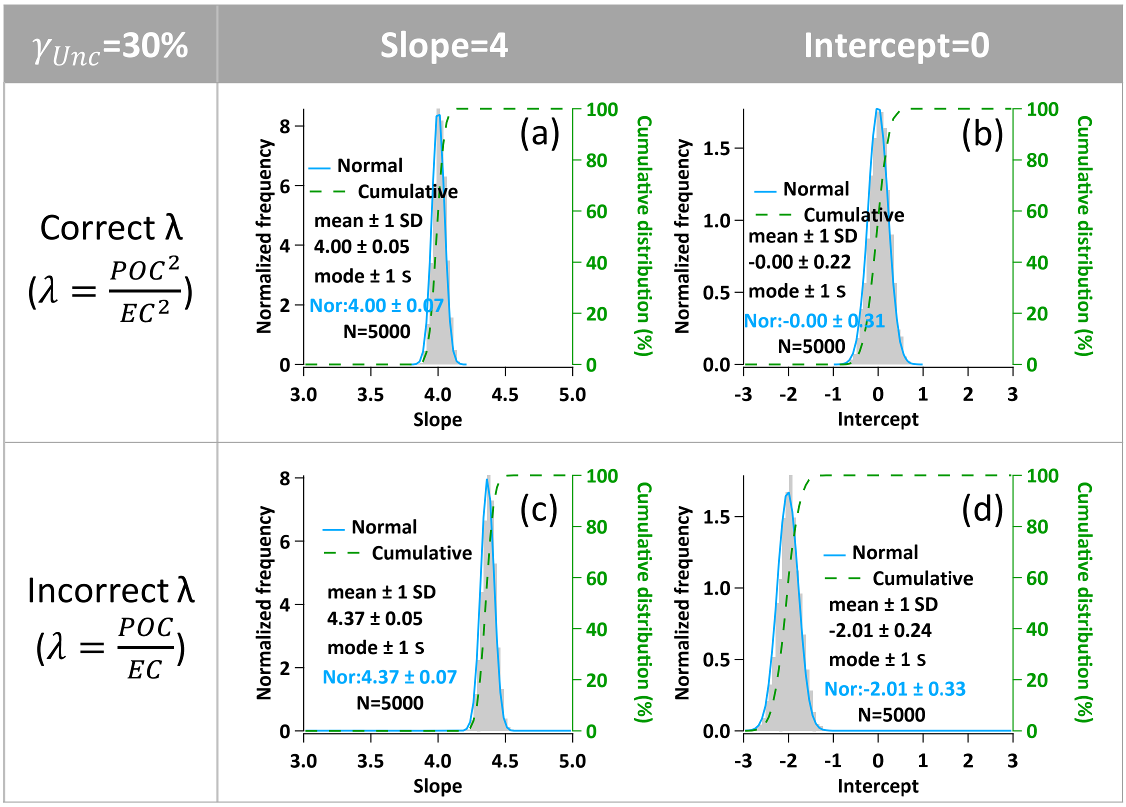


Figure S2. Example of bias in slope and intercept due to improper λ assignment. Data generation: Slope=4, Intercept=0; linear γ_{Unc} (30%). (a)&(b) Slopes and intercepts when proper λ is input following linear γ_{Unc} ($\lambda = \frac{POC^2}{EC^2}$); (c)&(d) Slopes and intercepts when improper λ is input following non-linear γ_{Unc} ($\lambda = \frac{POC}{EC}$).

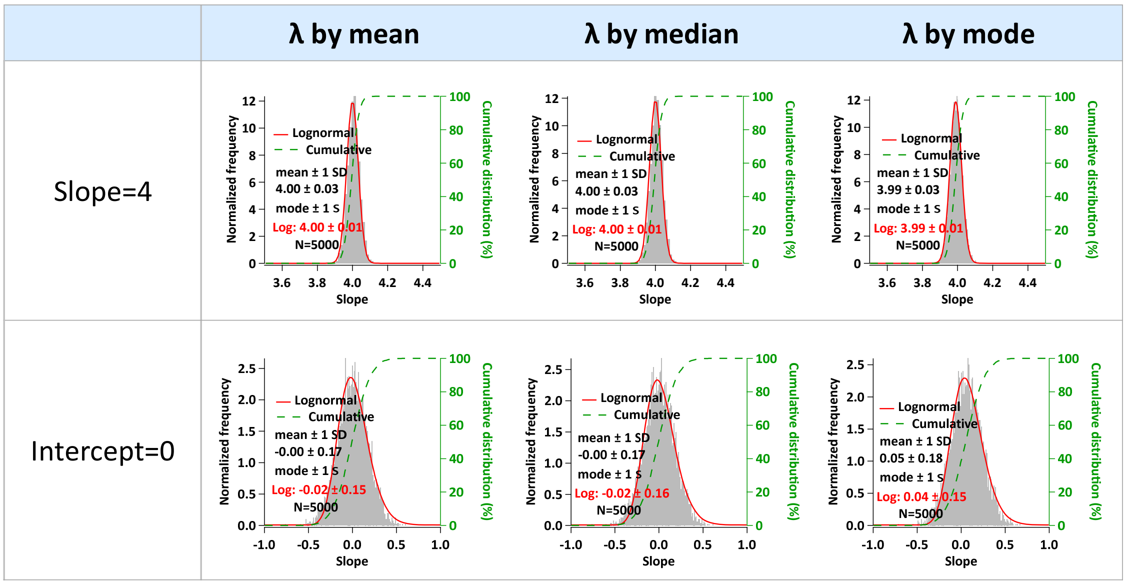


Figure S3. Sensitivity tests of λ calculated by mean, median and mode.

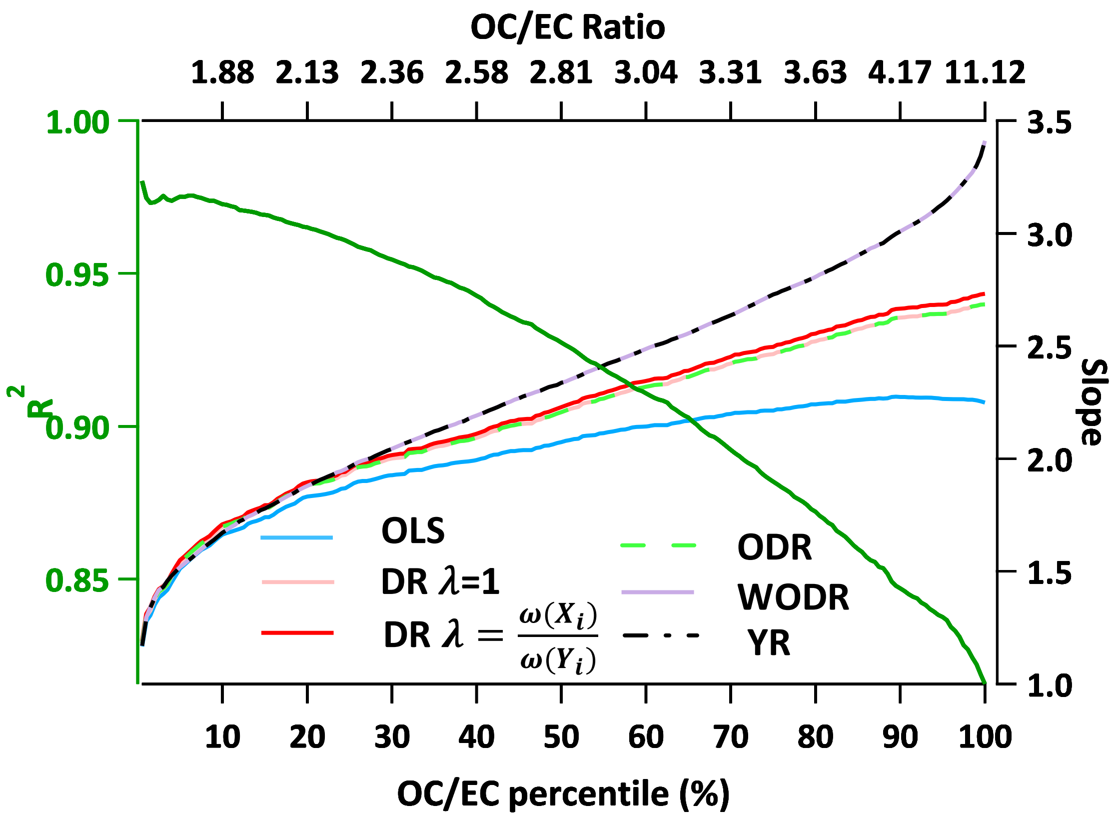


Figure S4. Regression slopes as a function of OC/EC percentile. OC/EC percentile range from 0.5% to 100%, with an interval of 0.5%.

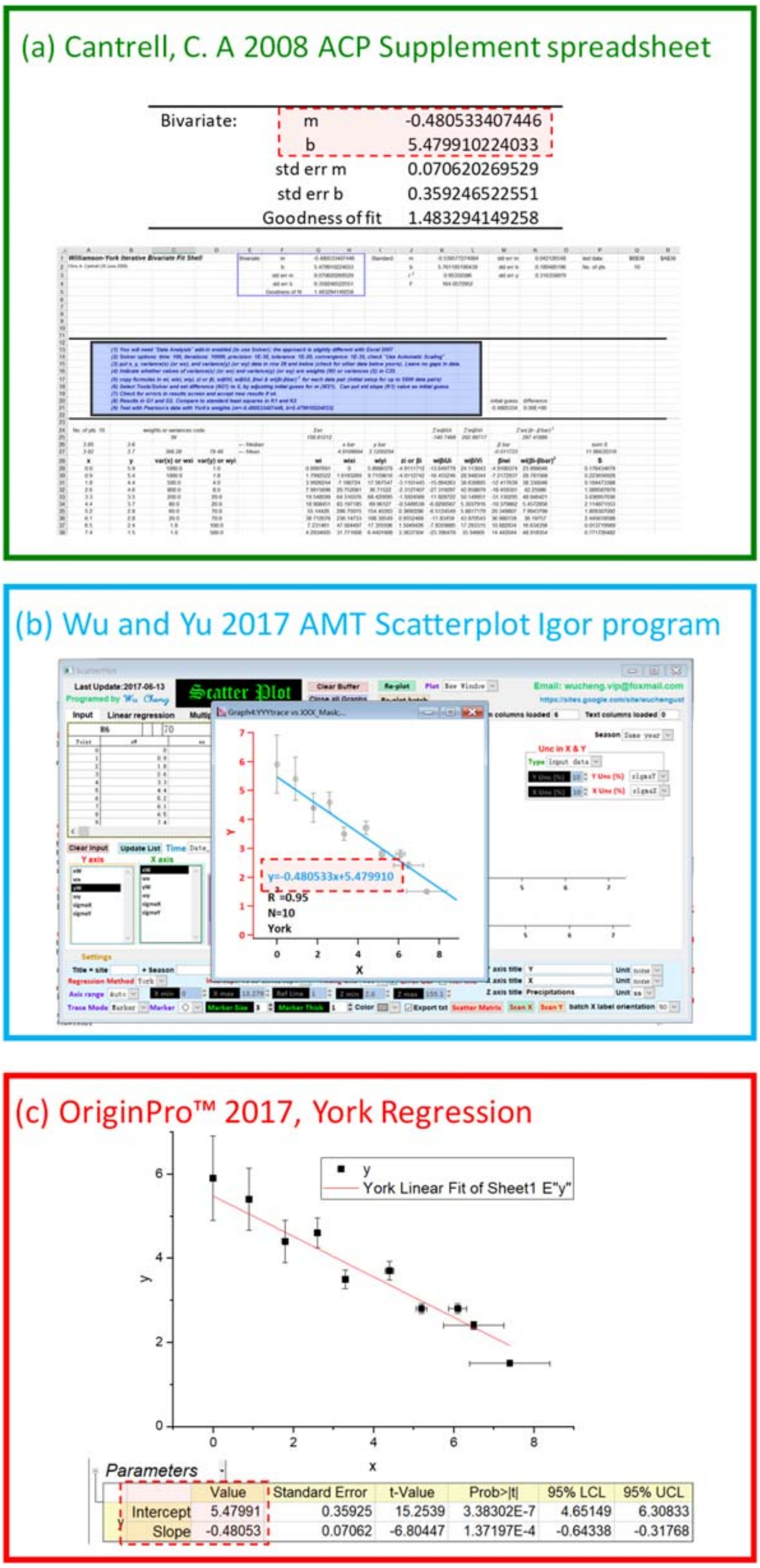


Figure S5. York regression implementations comparison, including spreadsheet by Cantrell (2008), Igor program by this study and a commercial software (OriginPro™ 2017).

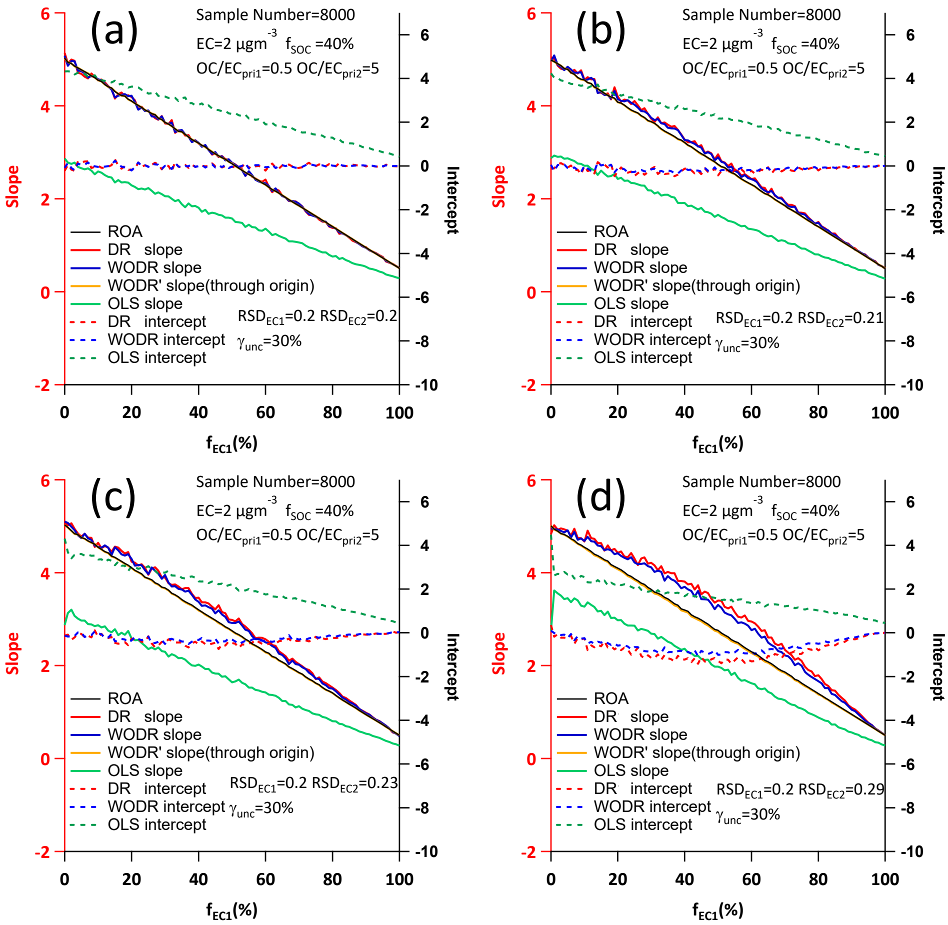


Figure S6. Study of two correlated sources secnario by different R^2 between the two sources. (a) $R^2 =$ 1 (b) $R^2 =$ 0.86 (c) $R^2 =$ 0.75 (d) $R^2 = 0.49$

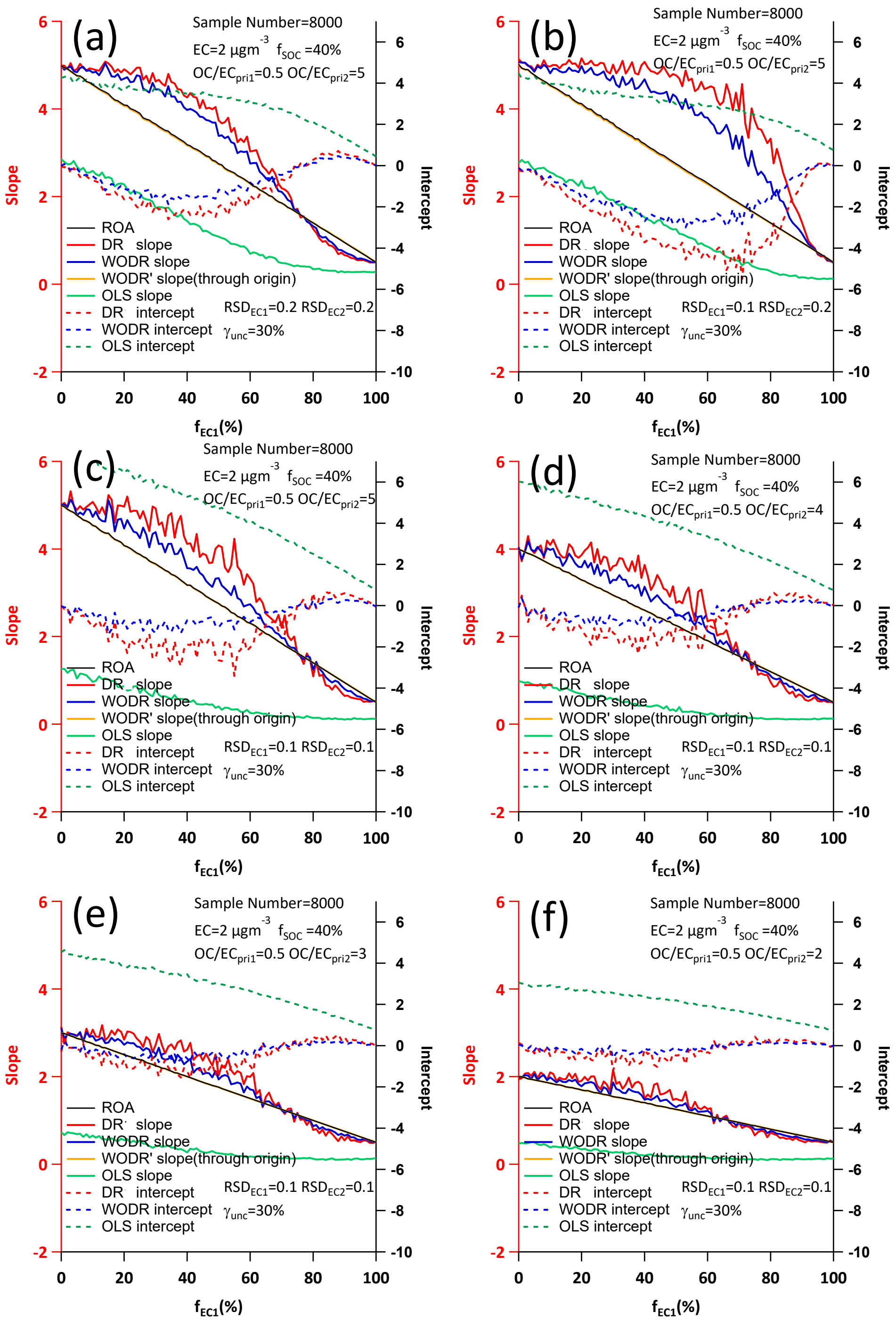


Figure S7. Study of two independent sources secnario by different parameters. (a)γ_pri=10, RSD$_{EC1}$=0.2,RSD$_{EC2}$=0.2 (b) γ_pri=10, RSD$_{EC1}$=0.1,RSD$_{EC2}$=0.2 (c) γ_pri=10, RSD$_{EC1}$=0.1, RSD$_{EC2}$=0.1 (d) γ_pri=8, RSD$_{EC1}$=0.1, RSD$_{EC2}$=0.1(e) γ_pri=6, RSD$_{EC1}$=0.1, RSD$_{EC2}$=0.1 (f) γ_pri=4, RSD$_{EC1}$=0.1, RSD$_{EC2}$=0.1

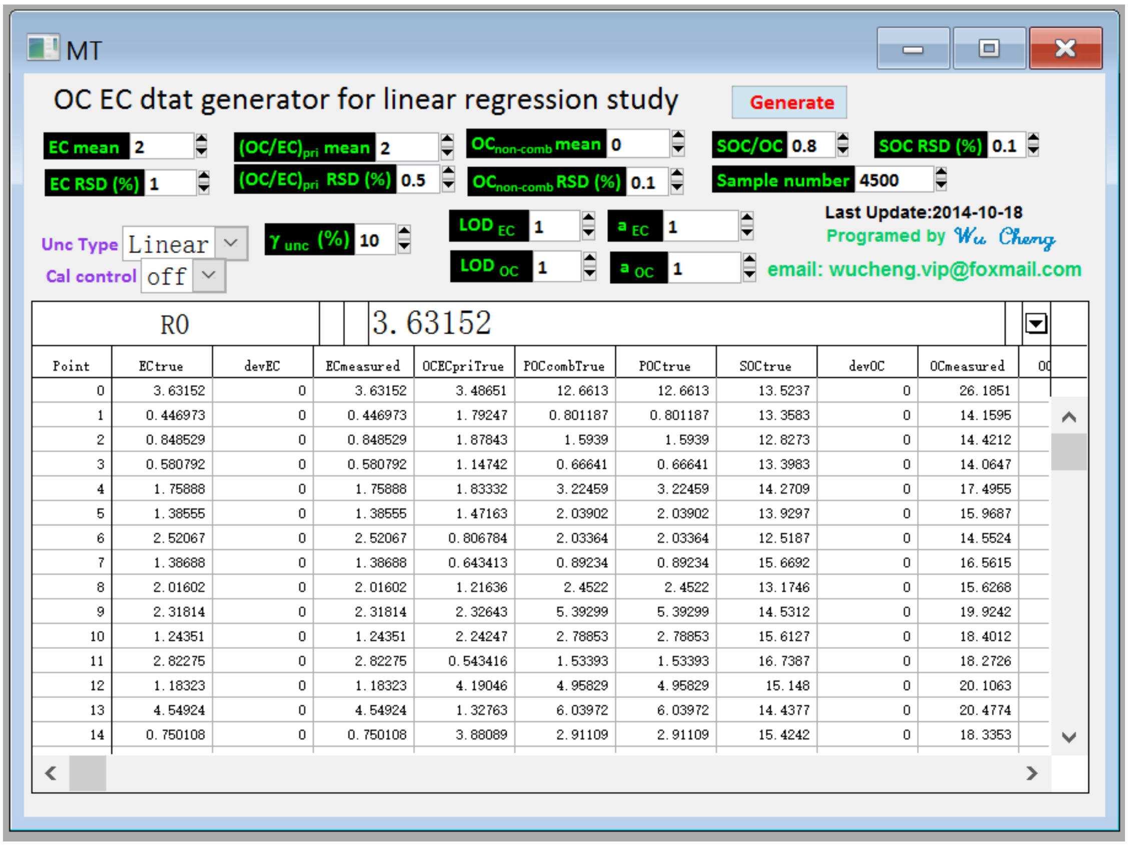


Figure S8. MT Igor program. OC and EC data following log-normal distribution can be generated for statistical study purpose (no time series information). User can define mean and RSD of EC, $(OC/EC)_{pri}$, SOC/OC ratio, measurement uncertainty, sample size, etc. MT Igor program can be downloaded from the following link: https://sites.google.com/site/wuchengust.