



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

Cheng Wu<sup>1,2</sup> and Jian Zhen Yu<sup>3,4,5</sup>

<sup>1</sup>Institute of Mass Spectrometer and Atmospheric Environment, Jinan University, Guangzhou 510632, China

<sup>2</sup>Guangdong Provincial Engineering Research Center for On-Line Source Apportionment System of Air Pollution, Guangzhou 510632, China

<sup>3</sup>Division of Environment, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China

<sup>4</sup>Atmospheric Research Centre, Fok Ying Tung Graduate School, Hong Kong University of Science and Technology, Nansha, China

<sup>5</sup>Department of Chemistry, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China

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

Received: 15 August 2017 – Discussion started: 27 September 2017

Revised: 6 February 2018 – Accepted: 6 February 2018 – Published: 2 March 2018

**Abstract.** Linear regression techniques are widely used in atmospheric science, but they 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 five techniques are ordinary least squares (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 nonlinear measurement uncertainties, (b) inclusion of a linear measurement uncertainty, and (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  $\lambda$  in DR is investigated by sensitivity tests, and it is found that an improper  $\lambda$  in DR can lead to a bias in both the slope and intercept estimation. Because the  $\lambda$  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 a priori error in one of the variables is unknown, or the measurement error described cannot be trusted, 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. An Igor Pro-based program (Scatter Plot) was developed to facilitate the implementation of error-in-variables regressions.

## 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; Lin et al., 2009), 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<sub>x</sub> 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 Bobruzki et al., 2010; Zieger et al., 2011; Wu et al., 2012; Huang et al., 2014; Zhou et al., 2016), inter-species analysis (Yu et al., 2005; Kuang et al., 2015), analytical protocol comparison (Chow et al., 2001, 2004; Cheng et al., 2011; Wu et al., 2016), light extinction budget reconstruction (Malm et al., 1994; Watson, 2002; Li et al., 2017), com-

parison 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<sub>2.5</sub> from MODIS data (Wang and Christopher, 2003), distinguishing OC origin from biomass burning using K<sup>+</sup> 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-variables 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 and thus recommended the use of DR for (OC/EC)<sub>pri</sub> (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 that 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 param-

eter,  $\lambda$ . This  $\lambda$  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-fit line for error-in-variables 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 variational 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 nonlinear 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 ( $\gamma_{\text{Unc}}$ ) is nonlinear with concentration, but a constant  $\gamma_{\text{Unc}}$  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 B1 for quick reference.

## 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 (i.e., distance of AB in Fig. S1 in the Supplement):

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

where  $Y_i$  are observed  $Y$  data points, while  $y_i$  are regressed  $Y$  data points of the regression line.  $N$  represents the number of data points that is used for regression.

### 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 (i.e., distance of AC in Fig. S1):

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

### Weighted orthogonal distance regression (WODR)

Unlike ODR, which considers even error in  $X$  and  $Y$ , weightings based on measurement errors in both  $X$  and  $Y$  are con-

sidered in WODR when minimizing the sum of squared orthogonal distance from the data points to the regression line (Carroll and Ruppert, 1996) as shown by AD in Fig. S1:

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

where  $\eta$  is the error variance ratio, which determines the angle  $\theta$  shown in Fig. S1. Implementation of ODR and WODR in Igor Pro (WaveMetrics, Inc. Lake Oswego, OR, USA) 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 as shown by AD in Fig. S1,

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

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  $Y_i$ ,

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

where  $\sigma_{X_i}$  and  $\sigma_{Y_i}$  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.

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

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.

Summary of the five regression techniques is given in Table S1 in the Supplement. It is worth noting that OLS and DR have closed-form expressions for calculating slope and intercept. In contrast, ODR, WODR and YR need to be solved iteratively. This need to be taken into consideration when choosing regression algorithm for handling huge numbers of data.

A computer program (Scatter Plot; Wu, 2017a) with a graphical user interface (GUI) in Igor Pro (WaveMetrics, Inc. Lake Oswego, OR, USA) was developed to facilitate the implementation of error-in-variables regression (including DR,

WODR and YR). Two other Igor Pro-based computer programs, Histbox (Wu, 2017b) and Aethalometer data processor (Wu, 2017c), are used for data analysis and visualization in this study.

## 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

$$Y = kX + b, \quad (7)$$

where  $k$  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:

$$\text{OC} = \text{POC}_{\text{comb}} + \text{POC}_{\text{non-comb}} + \text{SOC}, \quad (8)$$

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

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

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

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

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

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

The data generation starts from EC ( $X$  values). Once EC is generated,  $\text{POC}_{\text{comb}}$  (the part of  $Y$  that is correlated with  $X$ ) can be obtained by multiplying EC by a preset constant,  $(\text{OC}/\text{EC})_{\text{pri}}$  (slope  $k$ ). Then the other preset constant  $\text{POC}_{\text{non-comb}}$  is added to  $\text{POC}_{\text{comb}}$  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 is explained in Sect. 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 ( $\text{Conc.}_{\text{measured}}$ ) are simulated by adding measurement uncertainties ( $\varepsilon_{\text{Conc.}}$ ) to the true concentrations ( $\text{Conc.}_{\text{true}}$ ):

$$\text{Conc.}_{\text{measured}} = \text{Conc.}_{\text{true}} + \varepsilon_{\text{Conc.}}, \quad (12)$$

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

$$-\gamma_{\text{Unc}} \times \text{Conc.}_{\text{true}} \leq \varepsilon_{\text{Conc.}} \leq +\gamma_{\text{Unc}} \times \text{Conc.}_{\text{true}}. \quad (13)$$

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

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

$$\gamma_{\text{Unc-nonlinear}} = \frac{1}{\sqrt{\text{Conc.}_{\text{true}}}}, \quad (14)$$

and thus Eq. (13) for POC and EC becomes

$$-\frac{1}{\sqrt{\text{POC}_{\text{true}}}} \times \text{POC}_{\text{true}} \leq \varepsilon_{\text{POC}} \leq +\frac{1}{\sqrt{\text{POC}_{\text{true}}}} \times \text{POC}_{\text{true}}, \quad (15)$$

$$-\frac{1}{\sqrt{\text{EC}_{\text{true}}}} \times \text{EC}_{\text{true}} \leq \varepsilon_{\text{EC}} \leq +\frac{1}{\sqrt{\text{EC}_{\text{true}}}} \times \text{EC}_{\text{true}}. \quad (16)$$

In Eq. (14), the  $\gamma_{\text{Unc}}$  decreases as concentration increases, since low concentrations are usually more challenging to measure. As a result, the  $\gamma_{\text{Unc-nonlinear}}$  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  $\mu\text{g m}^{-3}$ , then the unit of  $\varepsilon_{\text{OC}}$  becomes  $\sqrt{\mu\text{g m}^{-3}}$ . Second, the concentration is not normalized by a consistent relative value, making it sensitive to the  $X$  and  $Y$  units used. For example, if  $\text{POC}_{\text{true}} = 0.9 \mu\text{g m}^{-3}$ , then  $\varepsilon_{\text{POC}} = \pm 0.95 \mu\text{g m}^{-3}$  and  $\gamma_{\text{Unc}} = 105\%$ , but by changing the concentration unit to  $\text{POC}_{\text{true}} = 900 \text{ ng m}^{-3}$ ,  $\varepsilon_{\text{OC}} = \pm 30 \text{ ng m}^{-3}$  and  $\gamma_{\text{Unc}} = 3\%$ . To overcome these deficiencies, we propose to modify Eq. (14) to

$$\gamma_{\text{Unc}} = \sqrt{\frac{\text{LOD}}{\text{Conc.}_{\text{true}}}} \times \alpha, \quad (17)$$

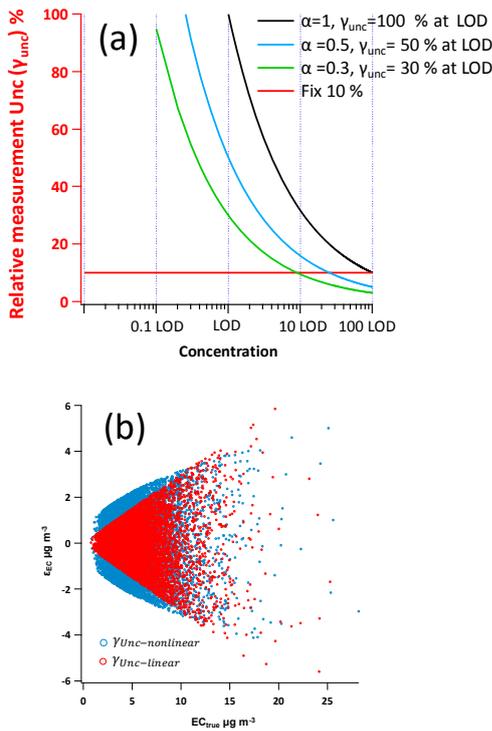
where LOD (limit of detection) is introduced to generate a dimensionless  $\gamma_{\text{Unc}}$ .  $\alpha$  is a dimensionless adjustable factor to control the position of  $\gamma_{\text{Unc}}$  curve on the concentration axis, which is indicated by the value of  $\gamma_{\text{Unc}}$  at LOD level. As shown in Fig. 1a, at different values of  $\alpha$  ( $\alpha = 1, 0.5$  and  $0.3$ ), the corresponding  $\gamma_{\text{Unc}}$  at the same LOD level would be 100, 50 and 30 %, respectively. By changing  $\alpha$ , the location of the  $\gamma_{\text{Unc}}$  curve on  $x$  axis direction can be set, using the  $\gamma_{\text{Unc}}$  at LOD as the reference point. Then Eq. (7) for POC and EC becomes

$$-\sqrt{\frac{\text{LOD}_{\text{POC}}}{\text{POC}_{\text{true}}}} \times \alpha_{\text{POC}} \times \text{POC}_{\text{true}} \leq \varepsilon_{\text{POC}} \leq +\sqrt{\frac{\text{LOD}_{\text{POC}}}{\text{POC}_{\text{true}}}} \times \alpha_{\text{POC}} \times \text{POC}_{\text{true}}, \quad (18)$$

$$-\sqrt{\frac{\text{LOD}_{\text{EC}}}{\text{EC}_{\text{true}}}} \times \alpha_{\text{EC}} \times \text{EC}_{\text{true}} \leq \varepsilon_{\text{EC}} \leq +\sqrt{\frac{\text{LOD}_{\text{EC}}}{\text{EC}_{\text{true}}}} \times \alpha_{\text{EC}} \times \text{EC}_{\text{true}}. \quad (19)$$

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

Uniform distribution has 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



**Figure 1.** (a) Example  $\gamma_{unc-nonlinear}$  curves by different  $\alpha$  values (Eq. 17). The  $x$  axis is concentration (normalized by LOD) in log scale and the  $y$  axis is  $\gamma_{unc}$ . Black, blue and green line represents  $\alpha$  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 %).

a uniform distribution in the interval  $[a, b]$ , the variance is  $\frac{1}{12}(a - b)^2$ . Since  $\varepsilon_{POC}$  and  $\varepsilon_{EC}$  follow a uniform distribution in the interval as given by Eqs. (18) and (19), the weights in DR and YR (inverse of variance) become

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

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

The parameter  $\lambda$  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:

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

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

where  $\gamma_{POC_{unc}}$  and  $\gamma_{EC_{unc}}$  are the relative measurement uncertainties, e.g., for relative measurement uncertainty of 10 %,  $\gamma_{unc} = 0.1$ . 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 Fig. 1b. For  $\gamma_{unc-linear}$ , the weights become

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

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

and  $\lambda$  for Deming regression can be determined:

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

### 3.1.2 XY data generation by the Mersenne twister 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 programming 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 lognormal) 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; and (4) unlike the sine function described by Chu (2005), which 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 Fig. 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 is considered in formulating the lognormal

distribution. The second step is to generate  $POC_{comb}$ . As shown in Fig. 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  $POC_{non-comb}$  onto  $POC_{comb}$ . Instead of having a distribution,  $POC_{non-comb}$  in this study is a single value, which favors direct comparison between the true value of  $POC_{non-comb}$  and  $POC_{non-comb}$  estimated from the regression intercept. The fourth step is to compute  $\epsilon_{POC}$  and  $\epsilon_{EC}$ . As discussed in Sect. 3.1.1, two types of measurement errors are considered for  $\epsilon_{POC}$  and  $\epsilon_{EC}$  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 a GUI was developed to facilitate the MT data generation for OC and EC. A brief introduction is given in the Supplement.

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

Besides MT, inclusion of the sine function data generation scheme in this study mainly serves two purposes. First, the sine function scheme was adopted in 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 yield the same results from the two previous studies. Second, the crosscheck between results from sine function and MT provides circumstantial evidence that the MT scheme works as expected.

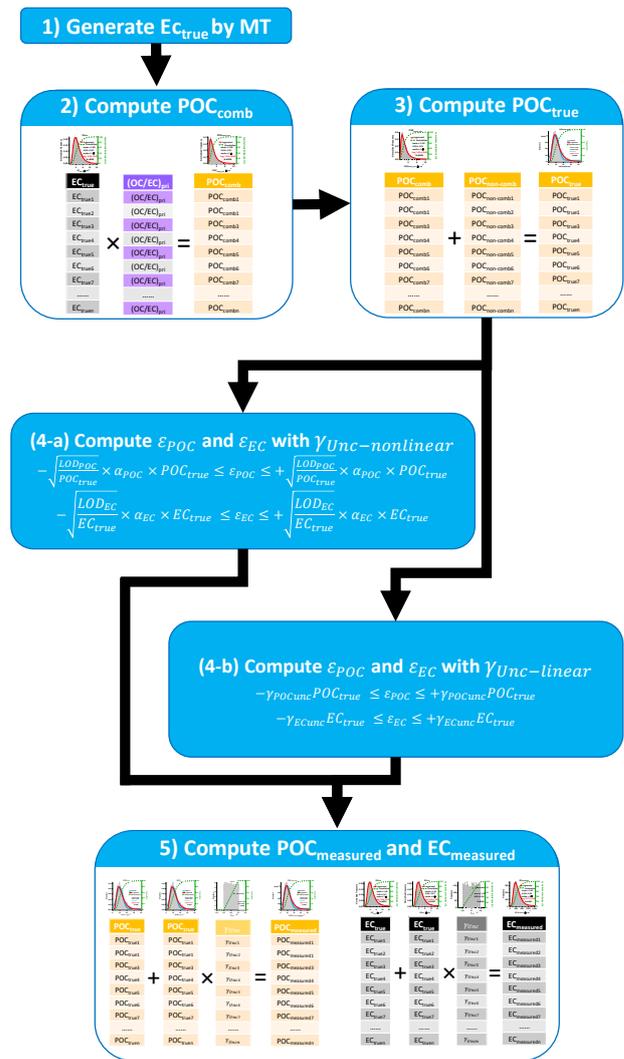
In this section, XY data generation by sine functions is demonstrated using POC as  $Y$  and EC as  $X$ . There are four steps in POC and EC data generation as shown by the flowchart in Fig. S2. Details are explained as follows. (1) The first step is to generate POC and EC (Chu, 2005):

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

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

where  $x$  is the elapsed hour ( $x = 1, 2, 3 \dots n; n \leq 120$ ),  $\tau$  is used to adjust the width of each peak, and  $\phi$  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 Fig. 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  $POC_{non-comb}$  to  $POC_{comb}$  to compute  $POC_{true}$ . As for  $POC_{non-comb}$ , a single value is assigned and added to all

**Data generation steps by MT**

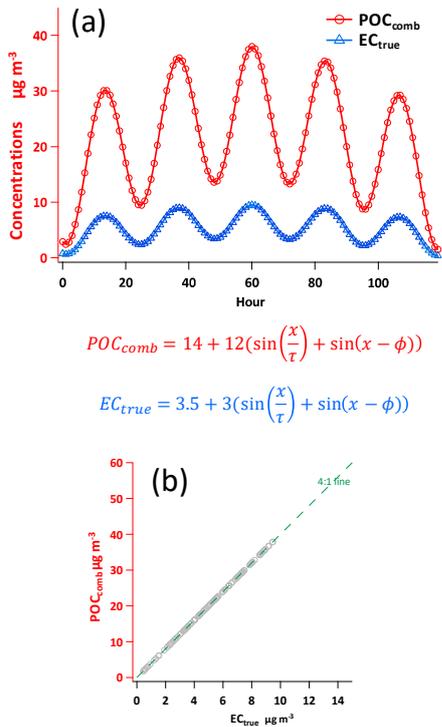


**Figure 2.** Flowchart of data generation steps using MT.

POC following Eq. (10). Then the goodness of the regression intercept can be evaluated by comparing the regressed intercept with preset  $POC_{non-comb}$ . (3) The third step is to compute  $\epsilon_{POC}$  and  $\epsilon_{EC}$ , 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  $\sigma_{abs}$  and EC**

Sampling was conducted from Feb 2012 to Jan 2013 at the suburban Nancun (NC) site ( $23^{\circ}0'11.82''N$ ,  $113^{\circ}21'18.04''E$ ), which is situated on top of the highest peak (141 m a.s.l.) in the Panyu district of Guangzhou. This site is located at the geographic center of Pearl River



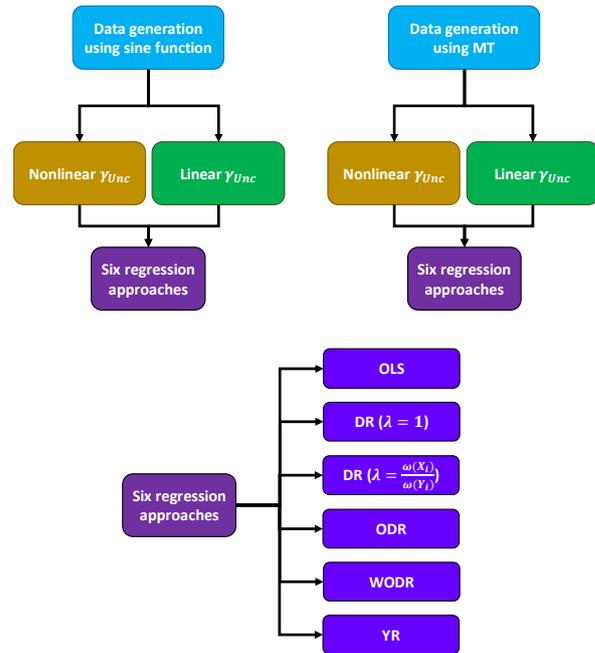
**Figure 3.**  $\text{POC}_{\text{comb}}$  and  $\text{EC}_{\text{true}}$  data generated by the sine functions of Chu (2005). (a) Time series of the 120 data points for  $\text{POC}_{\text{comb}}$  and  $\text{EC}_{\text{true}}$ . (b) Scatter plot of  $\text{POC}_{\text{comb}}$  vs.  $\text{EC}_{\text{true}}$ .

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  $7\lambda$  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  $2.5\mu\text{m}$  particle diameter cutoff. The algorithm of Weingartner et al. (2003) was adopted to correct the sampling artifacts (aerosol loading, filter matrix and scattering effect; Collaud Coen et al., 2010) in Aethalometer measurement. A customized computer program with GUI, Aethalometer data processor (Wu et al., 2018), 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. (2018).

#### 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 Fig. 4. Each data generation scheme considers both  $\gamma_{\text{Unc}}\text{-nonlinear}$  and  $\gamma_{\text{Unc}}\text{-linear}$  in measurement error parameterization. In total,

### Comparison study design



**Figure 4.** Overview of the comparison study design.

18 cases are tested with different combination of data generation schemes, measurement error parameterization schemes, true slope and intercept settings. In each case, six regression approaches are tested, i.e., OLS, DR ( $\lambda = 1$ ), DR ( $\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$ ), ODR, WODR and YR. In commercial software (e.g., OriginPro®, SigmaPlot®, GraphPad Prism®),  $\lambda$  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 dataset 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  $\gamma_{\text{Unc}}$ , six cases (Cases 1–6) are studied and the results are discussed below.

Table 1. Summary of six regression approaches comparison with 5000 runs for 18 cases.

Data generation				Results by different regression approaches													
Case	Data scheme	True Slope	True Intercept	$R^2$ (X, Y)	Measurement error	OLS		DR $\lambda = 1$		DR $\lambda = \frac{\sigma^2(X)}{\sigma^2(Y)}$		ODR		WODR		YR	
						slope	intercept	Slope	Intercept	Slope	Intercept	Slope	Intercept	Slope	Intercept	Slope	Intercept
1		4	0	0.67 ± 0.03	LOD <sub>PROC</sub> = 1, LOD <sub>DEC</sub> = 1, $\rho_{ROC} = 1$ , $\rho_{DEC} = 1$	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	$\rho_{ROC} = 1$ , $\rho_{DEC} = 1$ , $\rho_{ROC} = 1$ , $\rho_{DEC} = 1$	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	Chu	4	0	0.95 ± 0.01	LOD <sub>PROC</sub> = 0.5, LOD <sub>DEC</sub> = 0.5, $\rho_{ROC} = 0.5$ , $\rho_{DEC} = 0.5$	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 <sub>PROC</sub> = 1, LOD <sub>DEC</sub> = 0.5, $\rho_{ROC} = 1$ , $\rho_{DEC} = 1$	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	$\rho_{ROC} = 1$ , $\rho_{DEC} = 1$ , $\rho_{ROC} = 1$ , $\rho_{DEC} = 1$	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		4	0	0.76 ± 0.01		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	LOD <sub>PROC</sub> = 1, LOD <sub>DEC</sub> = 1, $\rho_{ROC} = 1$ , $\rho_{DEC} = 1$	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	$\rho_{ROC} = 1$ , $\rho_{DEC} = 1$ , $\rho_{ROC} = 1$ , $\rho_{DEC} = 1$	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	$\rho_{ROC} = 1$ , $\rho_{DEC} = 1$ , $\rho_{ROC} = 1$ , $\rho_{DEC} = 1$	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	MT	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	$\rho_{ROC} = 30\%$	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

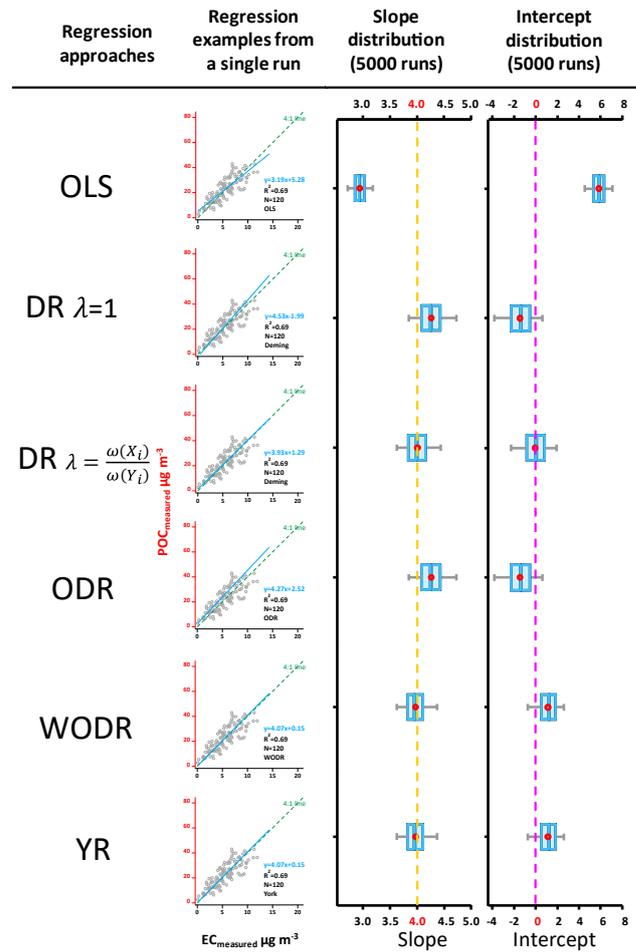
4.1.1 Results with  $\gamma_{\text{Unc-nonlinear}}$

A comparison of the regression techniques results with  $\gamma_{\text{Unc-nonlinear}}$  (following Eqs. 18 and 19) is summarized in Table 1.  $\text{LOD}_{\text{POC}}$ ,  $\text{LOD}_{\text{EC}}$ ,  $\alpha_{\text{POC}}$  and  $\alpha_{\text{EC}}$  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 nonzero intercept. The  $R^2$  (POC, EC) from 5000 runs for both Case 1 and 2 are  $0.67 \pm 0.03$ .

As shown in Fig. 5, for the zero-intercept case (Case 1), OLS significantly underestimates the slope ( $2.95 \pm 0.14$ ), while it overestimates the intercept ( $5.84 \pm 0.78$ ). 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  $\lambda$  is properly calculated by weights ( $\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$ ), unbiased slope ( $4.01 \pm 0.25$ ) and intercept ( $-0.04 \pm 1.28$ ) are obtained; however, results from DR with  $\lambda = 1$  show obvious bias in the slope ( $4.27 \pm 0.27$ ) and intercept ( $-1.45 \pm 1.36$ ). ODR also produces biased slope ( $4.27 \pm 0.27$ ) and intercept ( $-1.45 \pm 1.36$ ), which are identical to results of DR when  $\lambda = 1$ . With WODR, unbiased slope ( $3.98 \pm 0.22$ ) is observed, but the intercept is overestimated ( $1.12 \pm 1.02$ ). 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 1). The Case 2 intercepts are equal to the Case 1 intercepts plus 3, implying that all the regression methods are not sensitive to a nonzero intercept.

For Case 3,  $\text{LOD}_{\text{POC}} = 0.5$ ,  $\text{LOD}_{\text{EC}} = 0.5$ ,  $\alpha_{\text{POC}} = 0.5$ ,  $\alpha_{\text{EC}} = 0.5$  are adopted (Table 1), leading to an offset to the left of  $\gamma_{\text{Unc-nonlinear}}$  (blue curve) compared to Case 1 and 2 (black curve) in Fig. 1. As a result, for the same concentration of EC and OC in Case 3, the  $\gamma_{\text{Unc-nonlinear}}$  is smaller than in Cases 1 and 2 as indicated by a higher  $R^2$  ( $0.95 \pm 0.01$  for Case 3, Table 1). With a smaller measurement uncertainty, the degree of bias in Case 3 is smaller than in Case 1. For example, OLS slope is less biased in Case 3 ( $3.83 \pm 0.08$ ) compared to Case 1 ( $2.94 \pm 0.14$ ). Similarly, the slope ( $4.03 \pm 0.09$ ) and intercept ( $-0.18 \pm 0.44$ ) of DR ( $\lambda = 1$ ) 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.

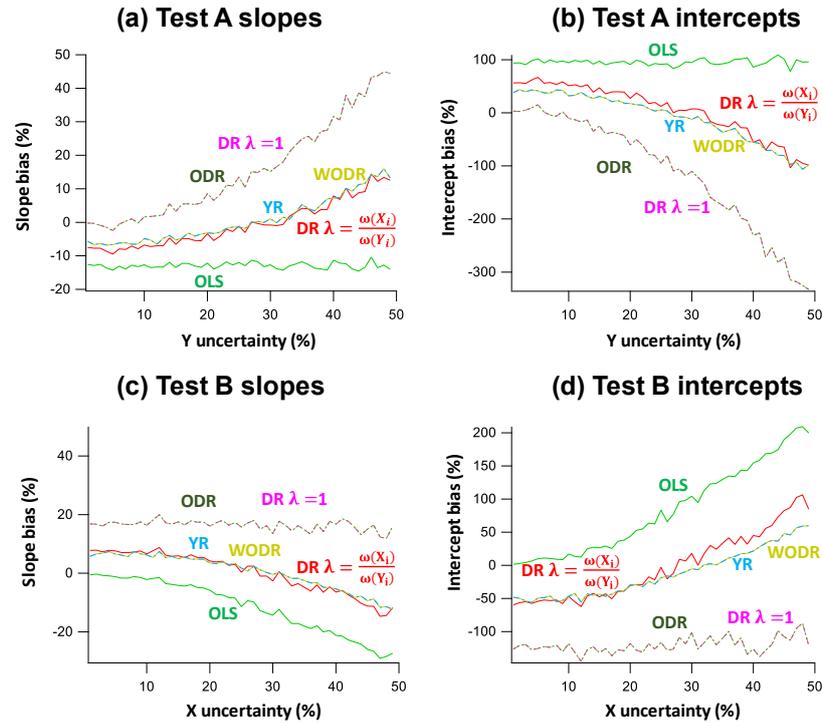
An uneven  $\text{LOD}_{\text{POC}}$  and  $\text{LOD}_{\text{EC}}$  is tested in Case 4 with  $\text{LOD}_{\text{POC}} = 1$ ,  $\text{LOD}_{\text{EC}} = 0.5$ ,  $\alpha_{\text{POC}} = 0.5$ ,  $\alpha_{\text{EC}} = 0.5$ , which yield an  $R^2$ (POC, EC) of  $0.78 \pm 0.02$ . The results are similar to 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.



**Figure 5.** Regression results on synthetic data, Case 1 (Slope = 4, Intercept = 0,  $\text{LOD}_{\text{POC}} = 1$ ,  $\text{LOD}_{\text{EC}} = 1$ ,  $\alpha_{\text{POC}} = 1$ ,  $\alpha_{\text{EC}} = 1$ ,  $R^2$  (POC, EC) =  $0.67 \pm 0.03$ ). 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 lines in orange and pink represent true slope and intercept, respectively.

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

Cases 5 and 6 represent the results from using  $\gamma_{\text{Unc-linear}}$  and are shown in Table 1.  $\gamma_{\text{Unc}}$  is set to 30% to achieve an  $R^2$  (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 ( $3.32 \pm 0.20$ ) and overestimates intercept ( $3.77 \pm 0.90$ ), while DR ( $\lambda = 1$ ) and ODR overestimate the slopes ( $4.75 \pm 0.30$ ) and underestimate the intercepts ( $-4.14 \pm 1.36$ ). In Case 6 (slope = 4, intercept = 3), results similar to Case 5 are obtained. It is worth noting that although the mean intercept ( $3.05 \pm 1.22$ ) of DR ( $\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$ )



**Figure 6.** Slope and intercept biases by different regression schemes in two test scenarios (A and B) in which the assumed error for one of the regression variables deviates from the actual measurement error. In Test A data generation,  $\gamma_{\text{Unc}_X}$  is fixed at 30% and  $\gamma_{\text{Unc}_Y}$  is varied between 1 and 50%. In Test B,  $\gamma_{\text{Unc}_X}$  varies between 1 and 50% and  $\gamma_{\text{Unc}_Y}$  is fixed at 30%. The assumed measurement error for regression is 10% for both X and Y. **(a)** Slope biases as a function of  $\gamma_{\text{Unc}_Y}$  in Test A. **(b)** Intercept biases as a function of  $\gamma_{\text{Unc}_Y}$  in Test A. **(c)** Slope biases as a function of  $\gamma_{\text{Unc}_X}$  in Test B. **(d)** Intercept biases as a function of  $\gamma_{\text{Unc}_X}$  in Test B.

is closest to the true value (intercept = 3), the deviations are much larger than for WODR ( $2.72 \pm 0.74$ ).

## 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  $\gamma_{\text{Unc-nonlinear}}$  and  $\gamma_{\text{Unc-linear}}$  are considered. With different configuration of slope, intercept and  $\gamma_{\text{Unc}}$ , 12 cases (Cases 7–18) are studied and the results are discussed below.

### 4.2.1 $\gamma_{\text{Unc-nonlinear}}$ results

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

DR ( $\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$ ) provides unbiased slope ( $4.00 \pm 0.03$ ) and intercept ( $3.00 \pm 0.18$ ) estimations. WODR and YR report unbiased slopes ( $3.97 \pm 0.03$ ) and overestimate intercepts ( $4.11 \pm 0.13$ ). 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,  $\text{LOD}_{\text{POC}} = 1$ ,  $\text{LOD}_{\text{EC}} = 1$ ,  $\alpha_{\text{POC}} = 1$ ,  $\alpha_{\text{EC}} = 1$ ) and Case 10 (slope = 0.5, intercept = 3,  $\text{LOD}_{\text{POC}} = 1$ ,  $\text{LOD}_{\text{EC}} = 1$ ,  $\alpha_{\text{POC}} = 1$ ,  $\alpha_{\text{EC}} = 1$ ) are conducted and the results are shown in Table 1. Unlike the overestimation observed in Cases 1–8, DR ( $\lambda = 1$ ) and ODR underestimate the slopes ( $0.46 \pm 0.01$ ) 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. Cases 11 and 12 test the bias when the true slope is 1, as shown in Table 1. 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 reports 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 $\gamma_{\text{Unc-linear}}$ results

Cases 13 and 14 (Table 1) represent the results from using  $\gamma_{\text{Unc-linear}}$  (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 ( $4.53 \pm 0.05$ ) and underestimate intercepts ( $-2.94 \pm 0.24$ ). 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 approach reporting unbiased intercept ( $3.08 \pm 0.23$ ). 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 1, the results are similar to Cases 13 and 14, i.e., 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 1, the results are similar to Cases 13 and 14, implying that these results are not sensitive to the special case when the true slope is 1.

#### 4.3 The importance of appropriate $\lambda$ input for Deming regression

As discussed above, inappropriate  $\lambda$  assignment in the Deming regression (e.g.,  $\lambda = 1$  by default for much commercial software) leads to biased slope and intercept. Besides  $\lambda = 1$ , inappropriate  $\lambda$  input due to improper handling of measurement uncertainty can also result in bias for Deming regression. An example is shown in Fig. S3. Data are generated by MT with following parameters: slope = 4, intercept = 0, and  $\gamma_{\text{Unc-linear}}$  (30 %). Figure S2a and b demonstrate that when an appropriate  $\lambda$  is provided (following  $\gamma_{\text{Unc-linear}}$ ,  $\lambda = \frac{\text{POC}^2}{\text{EC}^2}$ ), unbiased slopes and intercepts are obtained. If an improper  $\lambda$  is used due to a mismatched measurement uncertainty assumption ( $\gamma_{\text{Unc-nonlinear}}$ ,  $\lambda = \frac{\text{POC}}{\text{EC}}$ ), the slopes are overestimated (Fig. S3c,  $4.37 \pm 0.05$ ) and intercepts are underestimated (Fig. S3,  $-2.01 \pm 0.24$ ). This result emphasizes the importance of determining the correct form of measurement uncertainty in ambient samples, since  $\lambda$  is a crucial parameter in Deming regression.

In the  $\lambda$  calculation, different representations for POC and EC, including mean, median and mode, are tested as shown in Fig. S4. The results show that when  $X$  and  $Y$  have a similar distribution (e.g., both are lognormal), any of mean, median or mode can be used for the  $\lambda$  calculation.

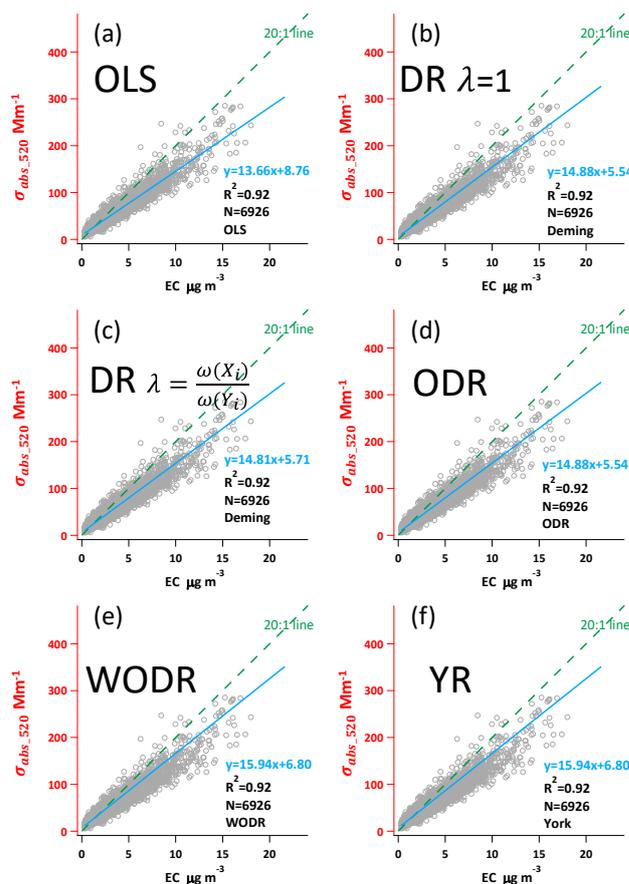


Figure 7. Regression results using ambient  $\sigma_{\text{abs}520}$  and EC data from a suburban site in Guangzhou, China.

#### 4.4 Caveats of regressions with unknown $X$ and $Y$ uncertainties

In atmospheric applications, there are scenarios in which a priori error in one of the variables is unknown, or the measurement error described cannot be trusted. For example, in the case of comparing model prediction and measurement data, the uncertainty of model prediction data is unknown. A second example is the case in which measurement uncertainty cannot be determined due to the lack of duplicated or collocated measurements and as a result, an arbitrarily assumed uncertainty is used. Such a case was illustrated in the study by Flanagan et al. (2006). They found that in the Speciation Trends Network (STN), the whole-system uncertainty retrieved by data from collocated samplers was different from the arbitrarily assumed 5 % uncertainty. Additionally, the discrepancy between the actual uncertainty obtained through collocated samplers and the arbitrarily assumed uncertainty varied by chemical species. To investigate the performance of different regression approaches in these cases, two tests (A and B) are conducted.

In Test A, the actual measurement error for  $X$  is fixed at 30 %, while  $\gamma_{\text{Unc}_Y}$  for  $Y$  varies from 1 to 50 %. The assumed measurement error for regression is 10 % for both  $X$  and  $Y$ . Results of Test A are shown in Fig. 6a and b. For OLS, the slopes are underestimated (−14 to −12 %) and intercepts are overestimated (90–103 %) and the biases are independent of variations in  $\gamma_{\text{Unc}_Y}$ . 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 the  $\gamma_{\text{Unc}_Y}$ . WODR, DR ( $\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$ ) and YR perform much better than other regression approaches in Test A, with a smaller bias in both slopes (−8–12 %) and intercepts −98–55 %.

In Test B,  $\gamma_{\text{Unc}_Y}$  is fixed at 30 % and  $\gamma_{\text{Unc}_X}$  varies between 1 and 50 %. The results of Test B are shown in Fig. 6c and d. The assumed measurement error for regression is 10 % for both  $X$  and  $Y$ . OLS underestimates the slopes (−29 to −0.2 %) and overestimates the intercepts (2–209 %). In contrast to Test A in which slope and intercept biases are independent of variations in  $\gamma_{\text{Unc}_Y}$ , the slope and intercept biases in Test B exhibit dependency on  $\gamma_{\text{Unc}_X}$ . The reason for this is that OLS only considers errors in  $Y$  and  $X$  is assumed to be error-free. ODR and DR ( $\lambda = 1$ ) yield similar results with overestimated slopes (11–18 %) and underestimated intercepts (−144 to −87 %). The degree of bias in slopes and intercepts is relatively independent on the  $\gamma_{\text{Unc}_X}$ . WODR, DR ( $\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$ ) and YR performed much better than the 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, if one of the measurement errors 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 and DR ( $\lambda = 1$ ) and OLS. This conclusion is consistent with results presented in Sect. 4.1 and 4.2. This analysis, albeit crude, also suggests that, in general, the magnitude of bias in slope estimation by these regression approaches is smaller than those for intercept. In other words, slope is a more reliable quantity compared to intercept when extracting quantitative information from linear regressions.

## 5 Regression applications to ambient data

This section demonstrates the application of the six regression approaches on a light absorption coefficient and EC dataset collected in a suburban site in Guangzhou. As mentioned in Sect. 4.4, measurement uncertainties are crucial inputs for DR, YR and WODR. The measurement precision of Aethalometer is 5 % (Hansen, 2005), while EC by the RT-ECOC analyzer is 24 % (Bauer et al., 2009). These measurement uncertainties are used in DR, YR and WODR calculation. The dataset contains 6926 data points with an  $R^2$  of 0.92.

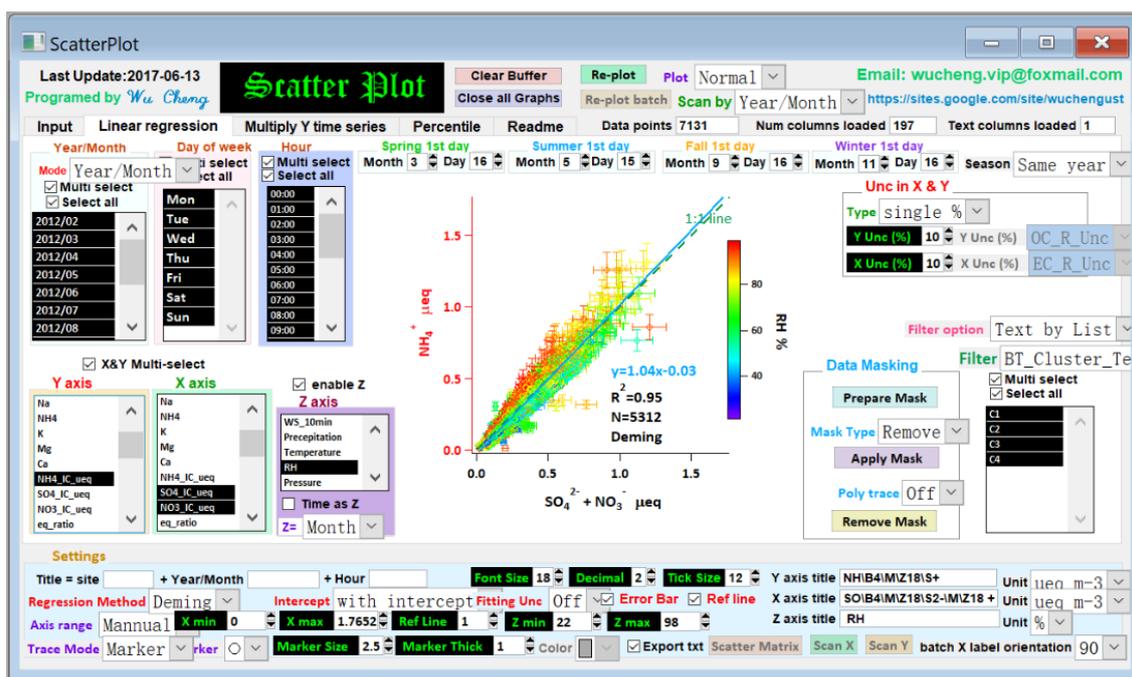
As shown in Fig. 7, the  $y$  axis is light absorption at 520 nm ( $\sigma_{\text{abs}520}$ ) and the  $x$  axis is EC mass concentration. The regressed slopes represent the mass absorption efficiency (MAE) of EC at 520 nm, ranging from 13.66 to 15.94 m<sup>2</sup> g<sup>−1</sup> by the six regression approaches. OLS yields the lowest slope (13.66 as shown in Fig. 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 S5 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 in 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 an  $R^2$  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, i.e., 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



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

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  $\lambda$  in DR. An improper  $\lambda$  could lead to biased slope and intercept. Since the  $\lambda$  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 a GUI in Igor Pro (WaveMetrics, Inc. Lake Oswego, OR, USA) was developed (Fig. 8). It is packed with many useful features for data analysis and plotting, including batch plotting, data masking via GUI, color coding in the  $z$  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>.

*Data availability.* OC, EC and  $\sigma_{\text{abs}}$  data used in this study are available from the corresponding authors upon request. The computer programs used for data analysis and visualization in this study are available in Wu (2017a–c).

### Appendix A: Equations of regression techniques

Ordinary least squares (OLS) calculation steps.

First calculate average of observed  $X_i$  and  $Y_i$ .

$$\bar{X} = \frac{\sum_{i=1}^N X_i}{N} \quad (\text{A1})$$

$$\bar{Y} = \frac{\sum_{i=1}^N Y_i}{N} \quad (\text{A2})$$

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

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

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

OLS slope and intercept can be obtained from

$$k = \frac{S_{yy}}{S_{xx}}, \quad (\text{A5})$$

$$b = \bar{Y} - k\bar{X}. \quad (\text{A6})$$

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

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

$$S_{xy} = \sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y}), \quad (\text{A7})$$

DR slope and intercept can be obtained from

$$k = \frac{S_{yy} - \lambda S_{xx} + \sqrt{(S_{yy} - \lambda S_{xx})^2 + 4\lambda S_{xy}^2}}{2S_{xy}}, \quad (\text{A8})$$

$$b = \bar{Y} - k\bar{X}. \quad (\text{A9})$$

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

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

$$W_i = \frac{\omega(X_i)\omega(Y_i)}{\omega(X_i) + k^2\omega(Y_i) - 2kr_i\sqrt{\omega(X_i)\omega(Y_i)}} \quad (\text{A10})$$

$$U_i = X_i - \bar{X} = X_i - \frac{\sum_{i=1}^N W_i X_i}{\sum_{i=1}^N W_i} \quad (\text{A11})$$

$$V_i = Y_i - \bar{Y} = Y_i - \frac{\sum_{i=1}^N W_i Y_i}{\sum_{i=1}^N W_i} \quad (\text{A12})$$

Then calculate  $\beta_i$ .

$$\beta_i = W_i \left[ \frac{U_i}{\omega(Y_i)} + \frac{kV_i}{\omega(X_i)} - [kU_i + V_i] \frac{r_i}{\sqrt{\omega(X_i)\omega(Y_i)}} \right] \quad (\text{A13})$$

Slope and intercept can be obtained from

$$k = \frac{\sum_{i=1}^N W_i \beta_i V_i}{\sum_{i=1}^N W_i \beta_i U_i}, \quad (\text{A14})$$

$$b = \bar{Y} - k\bar{X}. \quad (\text{A15})$$

Since  $W_i$  and  $\beta_i$  are functions of  $k$ ,  $k$  must be solved iteratively by repeating Eqs. (A11) to (A15). If the difference between the  $k$  obtained from Eq. (A15) and the  $k$  used in Eq. (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 dataset of Chu (2005) is around 6.

**Appendix B: Summary of abbreviations and symbols**

Abbreviation/symbol	Definition
$\alpha$	a dimensionless adjustable factor to control the position of $\gamma_{\text{Unc}}$ curve on the concentration axis
$b$	intercept in linear regression
$\beta_i, U_i, V_i, W_i$	intermediates in York regression calculations
$\gamma_{\text{Unc}}$	fractional measurement uncertainties relative to the true concentration (%)
DR	Deming regression
$\varepsilon_{\text{EC}}, \varepsilon_{\text{POC}}$	absolute measurement uncertainties of EC and POC
EC	elemental carbon
$\text{EC}_{\text{true}}$	numerically synthesized true EC concentration without measurement uncertainty
$\text{EC}_{\text{measured}}$	EC with measurement error ( $\text{EC}_{\text{true}} + \varepsilon_{\text{EC}}$ )
$\lambda$	$\omega(X_i)$ 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
$(\text{OC} / \text{EC})_{\text{pri}}$	primary OC / EC ratio
$\text{OC}_{\text{non-comb}}$	OC from non-combustion sources
ODR	orthogonal distance regression
OLS	ordinary least squares regression
POC	primary organic carbon
$\text{POC}_{\text{comb}}$	numerically synthesized true POC from combustion sources (well correlated with $\text{EC}_{\text{true}}$ ), measurement uncertainty not considered
$\text{POC}_{\text{non-comb}}$	numerically synthesized true POC from non-combustion sources (independent of $\text{EC}_{\text{true}}$ ) without considering measurement uncertainty
$\text{POC}_{\text{true}}$	sum of $\text{POC}_{\text{comb}}$ and $\text{POC}_{\text{non-comb}}$ without considering measurement uncertainty
$\text{POC}_{\text{measured}}$	POC with measurement error ( $\text{POC}_{\text{true}} + \varepsilon_{\text{POC}}$ )
$\sigma_{X_i}, \sigma_{Y_i}$	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
$\tau$	parameter in the sine function of Chu (2005) that adjusts the width of each peak
$\phi$	parameter in the sine function of Chu (2005) that adjusts the phase of the curve
WODR	weighted 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 $\sigma_{X_i}$ and $\sigma_{Y_i}$ , used as weights in DR calculation.

*Supplement.* The supplement related to this article is available online at: <https://doi.org/10.5194/amt-11-1233-2018-supplement>.

*Competing interests.* The authors declare that they have no conflict of interest.

*Acknowledgements.* This work was supported by the National Natural Science Foundation of China (grant no. 41605002, 41475004 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 Bin Yu Kuang at HKUST for the discussions on mathematics and Stephen M. Griffith at HKUST for the valuable comments.

Edited by: Willy Maenhaut

Reviewed by: two anonymous referees

## References

- Ayers, G. P.: Comment on regression analysis of air quality data, *Atmos. Environ.*, 35, 2423–2425, [https://doi.org/10.1016/S1352-2310\(00\)00527-6](https://doi.org/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.*, 59, 826–833, <https://doi.org/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 T. Math. Software*, 15, 348–364, <https://doi.org/10.1145/76909.76913>, 1989.
- Brauers, T. and Finlayson-Pitts, B. J.: Analysis of relative rate measurements, *Int. J. Chem. Kinet.*, 29, 665–672, [https://doi.org/10.1002/\(SICI\)1097-4601\(1997\)29:9<665::AID-KIN3>3.0.CO;2-S](https://doi.org/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, <https://doi.org/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, <https://doi.org/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., Kiehl, 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, <https://doi.org/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, <https://doi.org/10.1029/2000GL012354>, 2001.
- Cheng, Y., Duan, F.-K., He, K.-B., Zheng, M., Du, Z.-Y., Ma, Y.-L., and Tan, J.-H.: Intercomparison of thermal–optical methods for the determination of organic and elemental carbon: influences of aerosol composition and implications, *Environ. Sci. Technol.*, 45, 10117–10123, <https://doi.org/10.1021/es202649g>, 2011.
- Chow, J. C., Watson, J. G., Crow, D., Lowenthal, D. H., and Merrifield, T.: Comparison of IMPROVE and NIOSH carbon measurements, *Aerosol. Sci. Technol.*, 34, 23–34, <https://doi.org/10.1080/027868201300081923>, 2001.
- Chow, J. C., Watson, J. G., Chen, L. W. A., Arnott, W. P., and Moosmuller, H.: Equivalence of elemental carbon by thermal/optical reflectance and transmittance with different temperature protocols, *Environ. Sci. Technol.*, 38, 4414–4422, <https://doi.org/10.1021/Es034936u>, 2004.
- Chu, S. H.: Stable estimate of primary OC/EC ratios in the EC tracer method, *Atmos. Environ.*, 39, 1383–1392, <https://doi.org/10.1016/j.atmosenv.2004.11.038>, 2005.
- Collaud Coen, M., 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, <https://doi.org/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, *Meas. Tech.*, 46, 824–833, <https://doi.org/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, <https://doi.org/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, <https://doi.org/10.1016/j.atmosenv.2003.11.037>, 2004.
- Flanagan, J. B., Jayanty, R. K. M., Rickman, J. E. E., and Peterson, M. R.: PM<sub>2.5</sub> speciation trends network: evaluation of whole-system uncertainties using data from sites with collocated samplers, *J. Air Waste Manage.*, 56, 492–499, <https://doi.org/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 PM<sub>2.5</sub> major components and source investigation in suburban Hong Kong: A one year monitoring study, *Aerosol. Air. Qual. Res.*, 14, 237–250, <https://doi.org/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, <https://doi.org/10.5194/acp-10-1427-2010>, 2010.
- Kuang, B. Y., Lin, P., Huang, X. H. H., and Yu, J. Z.: Sources of humic-like substances in the Pearl River Delta, China: positive matrix factorization analysis of PM<sup>2.5</sup> major components and source markers, *Atmos. Chem. Phys.*, 15, 1995–2008, <https://doi.org/10.5194/acp-15-1995-2015>, 2015.
- Li, Y., Huang, H. X. H., Griffith, S. M., Wu, C., Lau, A. K. H., and Yu, J. Z.: Quantifying the relationship between visibility degradation and PM<sub>2.5</sub> constituents at a suburban site in Hong Kong: Differentiating contributions from hydrophilic and hydrophobic organic compounds, *Sci. Total Environ.*, 575, 1571–1581, <https://doi.org/10.1016/j.scitotenv.2016.10.082>, 2017.
- 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, <https://doi.org/10.1021/es990392d>, 1999.
- 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 PM<sub>2.5</sub> in Beijing and the estimation of secondary organic carbon, *J. Geophys. Res.*, 114, D00G11, <https://doi.org/10.1029/2008JD010902>, 2009.
- 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, <https://doi.org/10.1029/93JD02916>, 1994.
- Markovsky, I. and Van Huffel, S.: Overview of total least-squares methods, *Signal Process.*, 87, 2283–2302, <https://doi.org/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 T. Model. Comput. S.*, 8, 3–30, <https://doi.org/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, <https://doi.org/10.1029/98jd02618>, 1998.
- Petäjä, T., Mauldin III, 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, <https://doi.org/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, <https://doi.org/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, <https://doi.org/10.1016/j.atmosenv.2006.07.018>, 2006.
- Thompson, M.: Variation of precision with concentration in an analytical system, *Analyst*, 113, 1579–1587, <https://doi.org/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, [https://doi.org/10.1016/1352-2310\(94\)00276-Q](https://doi.org/10.1016/1352-2310(94)00276-Q), 1995.
- von Bobruzki, K., Braban, C. F., Famulari, D., Jones, S. K., Blackall, T., Smith, T. E. L., Blom, M., Coe, H., Gallagher, M., Ghaliyeni, 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, <https://doi.org/10.5194/amt-3-91-2010>, 2010.
- Wang, J. and Christopher, S. A.: Intercomparison between satellite-derived aerosol optical thickness and PM<sub>2.5</sub> mass: Implications for air quality studies, *Geophys. Res. Lett.*, 30, 2095, <https://doi.org/10.1029/2003gl018174>, 2003.
- Watson, J. G.: Visibility: Science and regulation, *J. Air Waste Manage.*, 52, 628–713, <https://doi.org/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, [https://doi.org/10.1016/S0021-8502\(03\)00359-8](https://doi.org/10.1016/S0021-8502(03)00359-8), 2003.
- Wu, C.: Scatter Plot, <https://doi.org/10.5281/zenodo.832417>, 2017a.
- Wu, C.: Aethalometer data processor, <https://doi.org/10.5281/zenodo.832404>, 2017b.
- Wu, C.: Histbox, <https://doi.org/10.5281/zenodo.832411>, 2017c.
- 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, <https://doi.org/10.5194/acp-16-5453-2016>, 2016.
- Wu, C., Ng, W. M., Huang, J., Wu, D., and Yu, J. Z.: Determination of Elemental and Organic Carbon in PM<sub>2.5</sub> in the Pearl River Delta Region: Inter-Instrument (Sunset vs. DRI Model 2001 Thermal/Optical Carbon Analyzer) and Inter-Protocol Comparisons (IMPROVE vs. ACE-Asia Protocol), *Aerosol. Sci. Technol.*, 46, 610–621, <https://doi.org/10.1080/02786826.2011.649313>, 2012.
- Wu, C., Huang, X. H. H., Ng, W. M., Griffith, S. M., and Yu, J. Z.: Inter-comparison of NIOSH and IMPROVE protocols for OC and EC determination: implications for inter-protocol data conversion, *Atmos. Meas. Tech.*, 9, 4547–4560, <https://doi.org/10.5194/amt-9-4547-2016>, 2016.
- Wu, C., Wu, D., and Yu, J. Z.: Quantifying black carbon light absorption enhancement with a novel statistical approach, *Atmos. Chem. Phys.*, 18, 289–309, <https://doi.org/10.5194/acp-18-289-2018>, 2018.
- York, D.: Least-squares fitting of a straight line, *Can. J. Phys.*, 44, 1079–1086, <https://doi.org/10.1139/p66-090>, 1966.
- York, D., Evensen, N. M., Martinez, 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, <https://doi.org/10.1119/1.1632486>, 2004.
- Yu, J. Z., Huang, X.-F., Xu, J., and Hu, M.: When Aerosol Sulfate Goes Up, So Does Oxalate: Implication for the Formation Mechanisms of Oxalate, *Environ. Sci. Technol.*, 39, 128–133, <https://doi.org/10.1021/es049559f>, 2005.

- 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, <https://doi.org/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 Roozendaal, 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, <https://doi.org/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 T. Math. Software*, 33, 27, <https://doi.org/10.1145/1268776.1268782>, 2007.