Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2017-300
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 27 September 2017

Atmospheric Measurement Techniques Discussions

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)

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2017-300
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 27 September 2017

Atmospheric Measurement Techniques Discussions

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 regression techniques 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. With the knowledge of an appropriate weighting, DR, WODR and YR are recommended for atmospheric studies when both x and y data have measurement errors.

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2017-300
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 27 September 2017

Atmospheric Measurement Techniques Discussions

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). Since the uncertainty of gravimetric analysis is much smaller compared to the uncertainty of the instrument response, the error free assumption in "x" is valid. However, in many other applications, such as inter-instrument comparison, x and y (from two instruments) may have a comparable degree of uncertainty. This deviation from the underlying assumption in OLS would produce biased slope and intercept when OLS is applied on the data set.

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, regression line should depends more on the precise data points rather than the less reliable data points. 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 ideal 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 is the key to handle the uneven errors between data points for the best fitting line calculation. This example demonstrates the importance of appropriate weighting in the calculation of fitting 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 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:

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

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2017-300
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 27 September 2017

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,

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

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.

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

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.

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 of the real-world application.

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2017-300
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 27 September 2017

Atmospheric Measurement Techniques Discussions

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:

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

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 relatively 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:

$$OC = POC_{comb} + POC_{non-comb} + SOC \tag{8}$$

Here POC_{comb} is primary OC from combustion. $POC_{non\text{-}comb}$ 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:

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

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

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

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

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

The data generation starts from EC (X values). Once EC is generated, POC_{comb} (correlated part of Y) can be obtained by multiplying EC with a preset constant, $(OC/EC)_{pri}$ (slope k). Then the other preset constant $POC_{non\text{-}comb}$ is added on POC_{comb} and the sum becomes POC (Y values). To simulate the real-world situation,

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2017-300
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 27 September 2017

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 ($\varepsilon_{Conc.}$) into the true concentrations ($Conc._{true}$):

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

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

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

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 $Conc._{true}$ 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 non-linear γ_{Unc}. In the data generation scheme of Chu (2005) for the measurement uncertainties (ε_{POC} and ε_{EC}), γ_{Unc} is non-linearly related to $Conc._{true}$:

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

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

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

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

The non-linear γ_{Unc} 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:

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2017-300
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 27 September 2017

Atmospheric Measurement Techniques Discussions

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

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:

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

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

With the improved non-linear γ_{Unc} 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.

Considering a uniform distribution of measurement error and defining the variance based on a single data point as was done above for the standard deviation, the weights in DR and YR (inverse of variance) are calculated from:

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

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

The parameter 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 non-linear γ_{Unc} discussed above, a second type measurement uncertainty with a constant γ_{Unc} is very common in atmospheric applications:

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

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

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2017-300
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 27 September 2017

where γ_{Unc} is the relative measurement uncertainty, 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 non-linear γ_{Unc} and linear γ_{Unc} is shown in Figure 1b. For a linear γ_{Unc}, 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 an 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

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2017-300
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 27 September 2017

Atmospheric Measurement Techniques Discussions

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: nonlinear γ_{Unc} and linear γ_{Unc}. 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 SI.

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

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 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 nonlinear γ_{Unc} and linear γ_{Unc}. (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.

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2017-300
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 27 September 2017

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 were equipped with a 2.5 μm inlet. 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 scenarios (RS) are compared using two data generation schemes (Sine function and MT) separately, as illustrated in Figure 4. Each data generation scheme considers both nonlinear γ_{Unc} and linear γ_{Unc} in measurement error parameterization. The six RS includes OLS, DR ($\lambda = 1$), DR ($\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$), ODR, WODR and YR. In commercial software (e.g., Origin, Sigma Pro, Prism, etc), λ in DR is sets 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 RS are considered to examine the sensitivity of regression results to different approaches for the various parameters. 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 6 regression scenarios. 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 nonlinear γ_{Unc}

A comparison of the regression techniques results with a non-linear γ_{Unc} (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.

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2017-300
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 27 September 2017

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 slopes (4.01±0.25) and intercepts (-0.04±1.28) are obtained, however, results from DR with λ=1 shows obvious bias in the slopes (4.27±0.27) and intercepts (-1.45±1.36). ODR also produces biased slopes (4.27±0.27) and intercepts (-1.45±1.36), which are identical to results of DR when λ=1. With WODR, unbiased slopes (3.98±0.22) are observed, but the intercepts are overestimated (1.12±1.02). Results of YR are identical to WODR. For Case 2 (slope=4, intercept=3), slopes from all six regression scenarios 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 S1), leading to an offset to the left of γ_{Unc} (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 γ_{Unc} is smaller than in Case 1 and Case 2 as indicated by higher the R^2 (0.95±0.01 for Case 3, Table S1). With a smaller measurement uncertainty, the degree of bias in Case 3 is smaller than Case 1. For example, OLS slopes are less biased in Case 3 (3.83±0.08) compare to Case 1 (2.94±0.14). Similarly, the slopes (4.03±0.09) and intercepts (-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 (Table S1) are similar to Case 1. For DR ($\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$) unbiased slopes and intercepts 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 linear γ_{Unc}

Cases 5 and 6 represent the results from using a constant γ_{Unc} and are shown in Table S2. 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 slopes (3.32 ±0.20) and overestimates intercepts (3.77 ±0.90), while

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2017-300
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 27 September 2017

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 6 regression scenarios. Both nonlinear γ_{Unc} and linear γ_{Unc} are considered. With different setup of slope, intercept and γ_{Unc}, 12 cases (Case 7 ~ Case 18) are studied and the results are discussed below.

4.2.1 Nonlinear γ_{Unc} results

Cases 7 and 8 use data generated by MT and a non-linear γ_{Unc} with results shown in Table 3. In Case 7 (slope=4, intercept=0, LOD_{POC}=1, LOD_{EC}=1, α_{POC}=1, α_{EC}=1), unbiased slopes (4.00 ±0.03) and intercepts (0.00 ±0.17) are 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 slopes (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 slopes (4.00 ±0.03) and intercepts (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 (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 S3. Unlike the overestimation observed in Case 1~Case 8, DR ($\lambda = 1$) and ODR underestimate the slope (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 S4. In Case 11 (intercept=0), all RS except OLS can provide unbiased results. In Case 12, all RS report unbiased slope except OLS, but DR ($\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$) is the only RS 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 slopes.

4.2.2 Fixed γ_{Unc} results

Cases 13 and 14 (results shown in Table S5) represent the results from using a linear γ_{Unc} (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 RS reports unbiased intercepts (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 S6, the results are similar to Table S5 that DR ($\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$) can provide unbiased slopes and intercepts 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 S7, the results are similar to those shown in Table S5, 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 linear γ_{Unc} (30%). Figure S2 a&b demonstrates that when an appropriate λ is provided (following linear γ_{Unc}, $\lambda = \frac{POC^2}{EC^2}$), unbiased slopes and intercepts are obtained. If an improper λ is used due to a mismatched measurement uncertainty assumption (non-linear γ_{Unc} , $\lambda = \frac{POC}{EC}$), 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.

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2017-300
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 27 September 2017

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 6, 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 scenarios. OLS yields the lowest slope (13.66 as shown in Figure 6a) among all six regression scenarios, consistent with the results by 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 RS exhibit a dependency on R^2. When percentile is relatively small (e.g., <10%), the difference between the six RS is also small due to the high R^2 (0.98). The deviations between the six RS become more pronounced as R^2 decrease (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 comparison results obtained 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. The results show that OLS fails to estimate the correct slope and intercept when

measurement errors are expected in both x and y, consistent with previous studies. With appropriate weighting, DR, WODR and YR are recommended for atmospheric studies when both the x and y data have measurement errors. 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 error during the experimentation stage rather than making assumptions. For ambient data with a XY R^2 less than 0.9, error-in-variables regression is needed to minimize the bias in slope and intercept. 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, a computer program (Scatter plot) with graphical user interface (GUI) in Igor Pro (WaveMetrics, Inc. Lake Oswego, OR, USA) is developed (Figure 7). It packed with lots of useful features for data analysis and graph 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.

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2017-300
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 27 September 2017

Atmospheric Measurement Techniques Discussions

Appendix A: Equations of regression techniques

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

First calculate average of observed X_i and Y_i.

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

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

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,

$$k = \frac{S_{yy}}{S_{xx}} \quad \text{(A6)}$$

$$b = \bar{Y} - k\bar{X} \quad \text{(A7)}$$

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

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

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

DR slope and intercept can be obtained from,

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

$$b = \bar{Y} - k\bar{X} \quad \text{(A10)}$$

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

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

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

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

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

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, the calculation is considered as converged. For the data set of Chu (2005), the iteration count is around 6.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (41605002). 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.

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.

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

496 Huang, X. H., Bian, Q., Ng, W. M., Louie, P. K., and Yu, J. Z.: Characterization of PM2. 5 major components
497 and source investigation in suburban Hong Kong: A one year monitoring study, Aerosol. Air. Qual. Res., 14,
498 237-250, doi: 10.4209/aaqr.2013.01.0020, 2014.

499 Janhäll, S., Andreae, M. O., and Pöschl, U.: Biomass burning aerosol emissions from vegetation fires: particle
500 number and mass emission factors and size distributions, Atmos. Chem. Phys., 10, 1427-1439, doi:
501 10.5194/acp-10-1427-2010, 2010.

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

504 Linnet, K.: Necessary sample size for method comparison studies based on regression analysis, Clin. Chem.,
505 45, 882-894, 1999.

506 Malm, W. C., Sisler, J. F., Huffman, D., Eldred, R. A., and Cahill, T. A.: Spatial and seasonal trends in particle
507 concentration and optical extinction in the United-States, J. Geophys. Res., 99, 1347-1370, doi:
508 10.1029/93JD02916, 1994.

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

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

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

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

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

521 Saylor, R. D., Edgerton, E. S., and Hartsell, B. E.: Linear regression techniques for use in the EC tracer method
522 of secondary organic aerosol estimation, Atmos. Environ., 40, 7546-7556, doi:
523 10.1016/j.atmosenv.2006.07.018, 2006.

524 Turpin, B. J. and Huntzicker, J. J.: Identification of secondary organic aerosol episodes and quantitation of
525 primary and secondary organic aerosol concentrations during SCAQS, Atmos. Environ., 29, 3527-3544, doi:
526 10.1016/1352-2310(94)00276-Q, 1995.

527 von Bobrutzki, K., Braban, C. F., Famulari, D., Jones, S. K., Blackall, T., Smith, T. E. L., Blom, M., Coe, H.,
528 Gallagher, M., Ghalaieny, M., McGillen, M. R., Percival, C. J., Whitehead, J. D., Ellis, R., Murphy, J.,
529 Mohacsi, A., Pogany, A., Junninen, H., Rantanen, S., Sutton, M. A., and Nemitz, E.: Field inter-comparison
530 of eleven atmospheric ammonia measurement techniques, Atmos. Meas. Tech., 3, 91-112, doi: 10.5194/amt-
531 3-91-2010, 2010.

532 Wang, J. and Christopher, S. A.: Intercomparison between satellite-derived aerosol optical thickness and
533 PM2.5 mass: Implications for air quality studies, Geophys. Res. Lett., 30, 2095, doi: 10.1029/2003gl018174,
534 2003.

535 Watson, J. G.: Visibility: Science and regulation, J. Air Waste Manage. Assoc., 52, 628-713, doi:
536 10.1080/10473289.2002.10470813, 2002.

537 Weingartner, E., Saathoff, H., Schnaiter, M., Streit, N., Bitnar, B., and Baltensperger, U.: Absorption of light
538 by soot particles: determination of the absorption coefficient by means of aethalometers, J. Aerosol. Sci., 34,
539 1445-1463, doi: 10.1016/S0021-8502(03)00359-8, 2003.

540 Wu, C. and Yu, J. Z.: Determination of primary combustion source organic carbon-to-elemental carbon
541 (OC/EC) ratio using ambient OC and EC measurements: secondary OC-EC correlation minimization method,
542 Atmos. Chem. Phys., 16, 5453-5465, doi: 10.5194/acp-16-5453-2016, 2016.

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

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

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

548 Zhou, Y., Huang, X. H. H., Griffith, S. M., Li, M., Li, L., Zhou, Z., Wu, C., Meng, J., Chan, C. K., Louie, P.
549 K. K., and Yu, J. Z.: A field measurement based scaling approach for quantification of major ions, organic
550 carbon, and elemental carbon using a single particle aerosol mass spectrometer, Atmos. Environ., 143, 300-
551 312, doi: 10.1016/j.atmosenv.2016.08.054, 2016.

552 Zieger, P., Weingartner, E., Henzing, J., Moerman, M., de Leeuw, G., Mikkilä, J., Ehn, M., Petäjä, T., Clémer,
553 K., van Roozendael, M., Yilmaz, S., Frieß, U., Irie, H., Wagner, T., Shaiganfar, R., Beirle, S., Apituley, A.,
554 Wilson, K., and Baltensperger, U.: Comparison of ambient aerosol extinction coefficients obtained from in-
555 situ, MAX-DOAS and LIDAR measurements at Cabauw, Atmos. Chem. Phys., 11, 2603-2624, doi:
556 10.5194/acp-11-2603-2011, 2011.

557 Zwolak, J. W., Boggs, P. T., and Watson, L. T.: Algorithm 869: ODRPACK95: A weighted orthogonal
558 distance regression code with bound constraints, ACM Trans. Math. Softw., 33, 27, doi:
559 10.1145/1268776.1268782, 2007.

560

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2017-300
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 27 September 2017

Atmospheric Measurement Techniques Discussions

561 **Table 1.** Abbreviations.

Abbreviation	Definition
α	a dimensionless adjustable factor to control the position of γ_{Unc} curve on the concentration axis
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
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
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
YR	York regression
$\omega(X_i)$, $\omega(Y_i)$	inverse of σ_{X_i} and σ_{Y_i}, used as weights in DR calculation.

562

563

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2017-300
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 27 September 2017

Atmospheric Measurement Techniques Discussions

564 **Table 2.** Comparison of regression techniques using the data generation scheme of Chu (2005) with 5000 runs
565 considering both zero intercept and non-zero intercept conditions. Non-linear relative measurement
566 uncertainty follows Eqs. (18) & (19). Unbiased (difference <5% to true value) slopes and intercepts are
567 highlighted in grey.

Case	Regression scenario	Slope mean	Slope SD	Intercept mean	Intercept SD
Case 1	OLS	2.94	±0.14	5.84	±0.78
Slope=4, Intercept=0	DR λ=1	4.27	±0.27	-1.45	±1.36
LOD_{POC}=1, LOD_{EC}=1 a_{POC}=1, a_{EC}=1. R²(POC,EC) = 0.67±0.03	DR $\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$	4.01	±0.25	-0.04	±1.28
	ODR	4.27	±0.27	-1.45	±1.36
	WODR	3.98	±0.22	1.12	±1.02
	YR	3.98	±0.22	1.12	±1.02
Case 2	OLS	2.95	±0.15	8.83	±0.80
Slope=4, Intercept=3	DR1 λ=1	4.32	±0.28	1.28	±1.43
LOD_{POC}=1, LOD_{EC}=1 a_{POC}=1, a_{EC}=1. R²(POC,EC) = 0.67±0.04	DR1 $\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$	4.01	±0.26	2.94	±1.34
	ODR	4.32	±0.28	1.28	±1.43
	WODR	3.99	±0.23	3.98	±1.05
	YR	3.99	±0.23	3.98	±1.05

568

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2017-300
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 27 September 2017

Atmospheric Measurement Techniques Discussions

569 **Table 3.** Comparison of regression techniques using MT data generation scheme with 5000 runs with non-
570 linear relative measurement uncertainty following Eqs. (18) & (19). Unbiased (difference <5% to true value)
571 slopes and intercepts are highlighted with grey background.

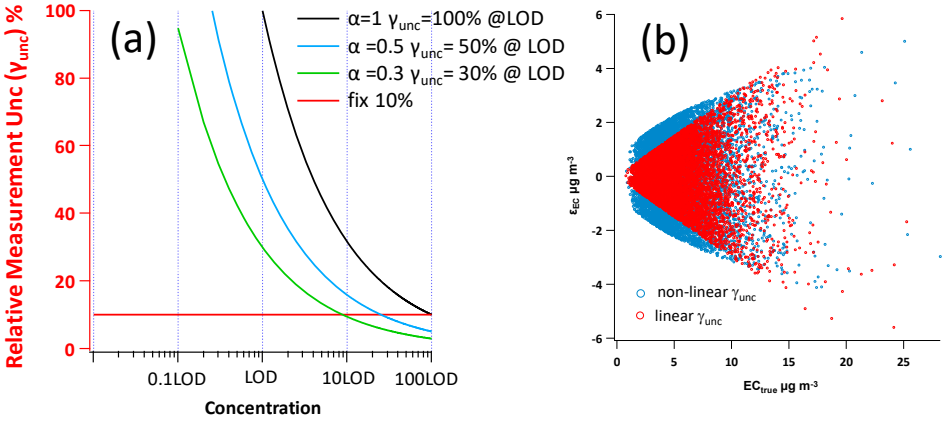
Case	Regression scenario	Slope mean	Slope SD	Intercept mean	Intercept SD
Case 7	OLS	3.22	±0.03	4.30	±0.14
Slope=4, Intercept=0	DR λ=1	4.17	±0.04	-0.94	±0.18
LOD_{POC}=1, LOD_{EC}=1 a_{POC}=1, a_{EC}=1.	DR $\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$	4.00	±0.03	0.00	±0.17
	ODR	4.17	±0.04	-0.94	±0.18
R^2 (POC,EC) = 0.76±0.01	WODR	3.96	±0.03	1.21	±0.13
	YR	3.96	±0.03	1.21	±0.13
Case 8	OLS	3.22	±0.03	7.29	±0.14
Slope=4, Intercept=3	DR λ=1	4.20	±0.04	1.88	±0.18
LOD_{POC}=1, LOD_{EC}=1 a_{POC}=1, a_{EC}=1.	DR $\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$	4.00	±0.03	3.00	±0.18
	ODR	4.20	±0.04	1.88	±0.18
R^2 (POC,EC) = 0.75±0.01	WODR	3.97	±0.03	4.11	±0.13
	YR	3.97	±0.03	4.11	±0.13

572

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2017-300
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 27 September 2017

Atmospheric Measurement Techniques Discussions

573



574

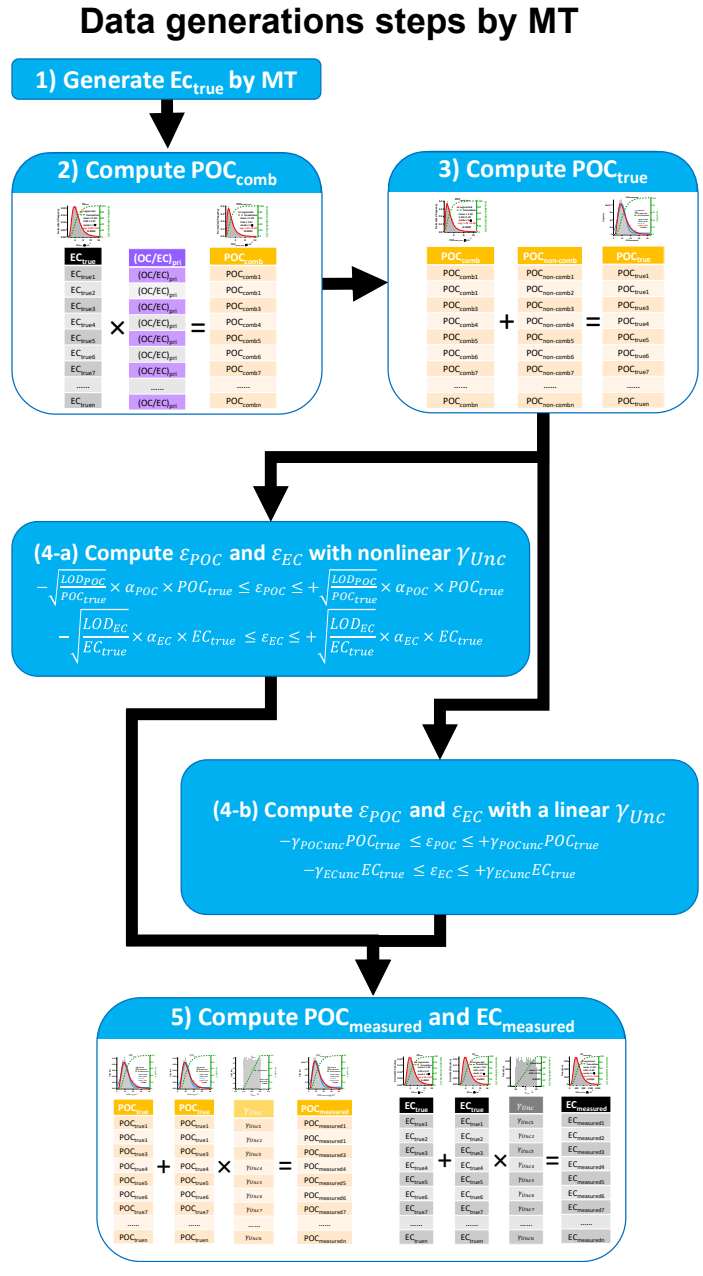
575 **Figure 1.** (a) Example γ_{Unc} curves by different α values (Eq. (17)). The x-axis is concentration (normalized
576 by LOD) and y axis is γ_{Unc}. Black, blue and green line represent α equal to 1, 0.5 and 0.3 respectively,
577 corresponding to the γ_{unc} at LOD level equals to 100%, 50% and 30% respectively. (b) Example of
578 measurement uncertainty generation with non-linear γ_{Unc} and linear γ_{Unc} constrain. The blue circles represent
579 non-linear γ_{Unc} following Eq. (17) ($LOD_{EC} = 1$, $a_{EC} = 1$). The red circles represent linear γ_{Unc} (30%).

580

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2017-300
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 27 September 2017

Atmospheric Measurement Techniques Discussions

581

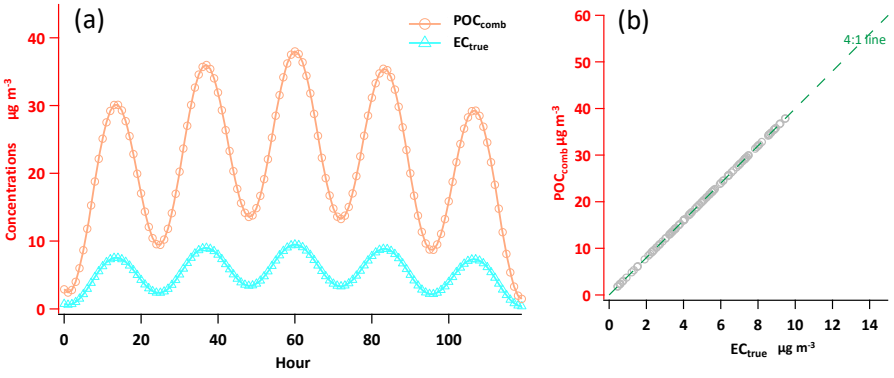


582

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

584

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2017-300
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 27 September 2017



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

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

585

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

588

589

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2017-300
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 27 September 2017

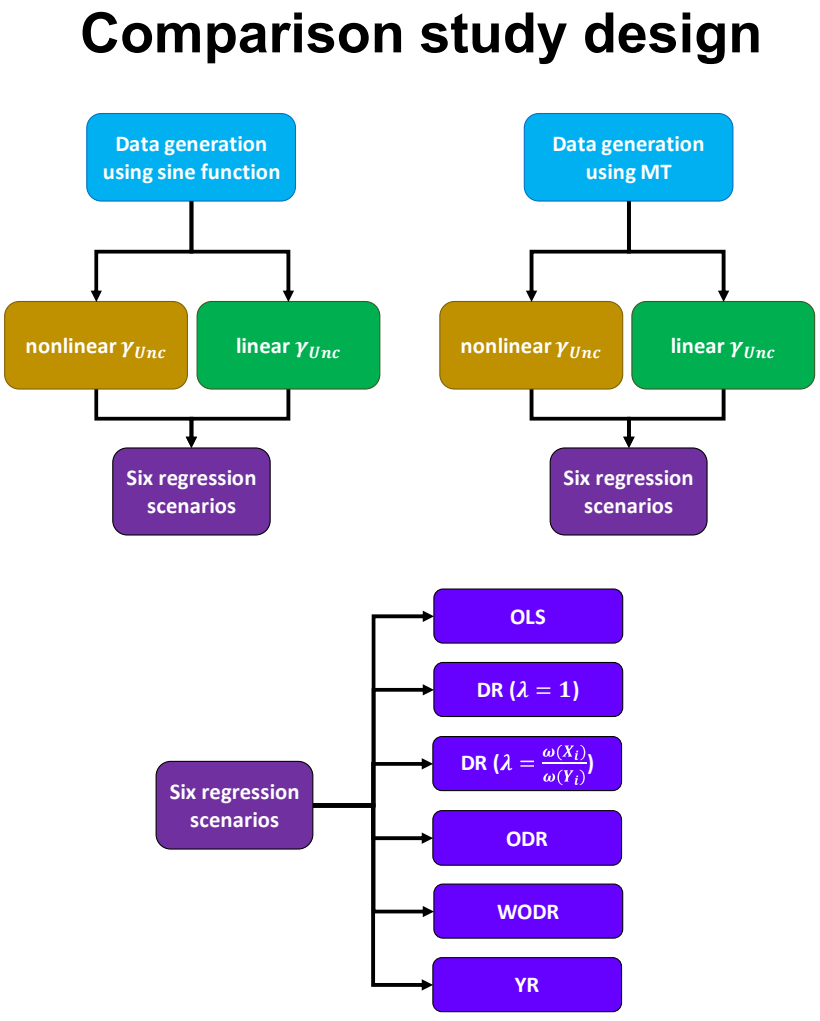


Figure 4. Overview of the comparison study design.

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2017-300
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 27 September 2017

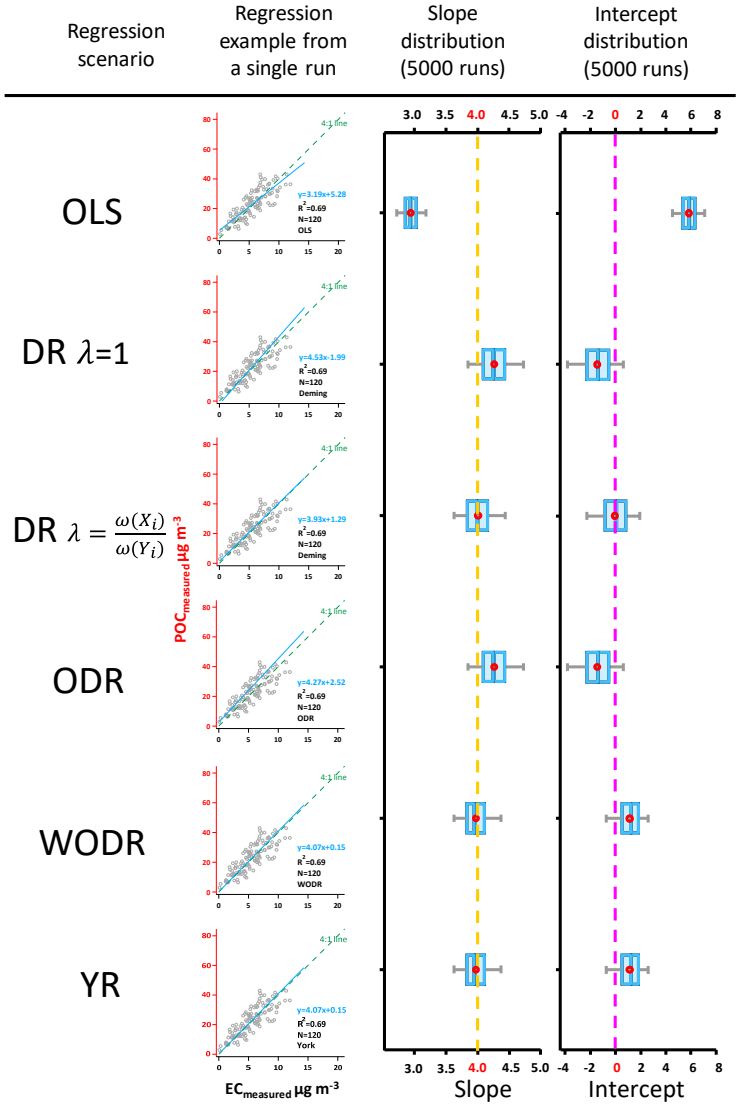
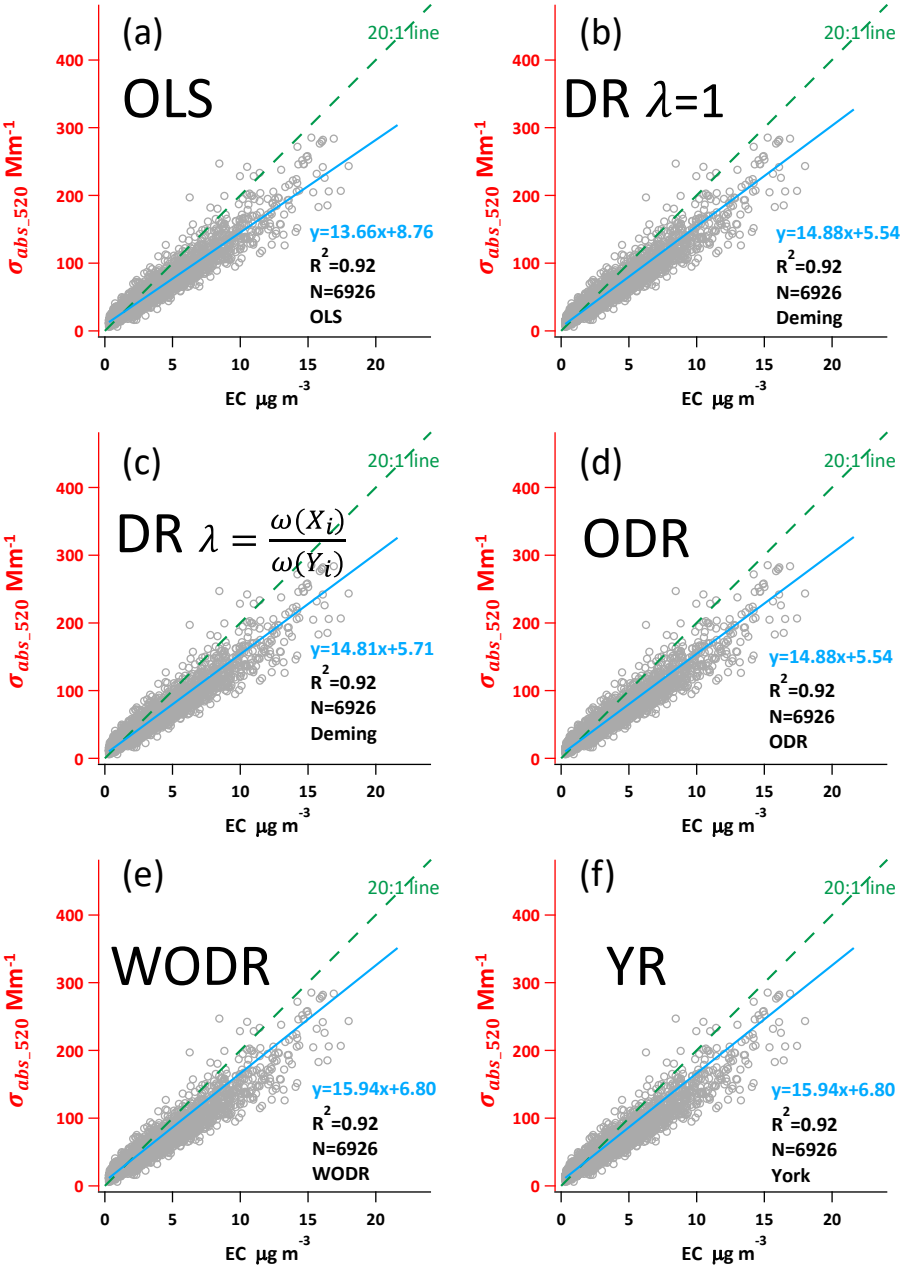


Figure 5. Regression results on synthetic data, case 1 (Slope=4, Intercept=0, LOD_{POC}=1, LOD_{EC}=1, a_{POC}=1, a_{EC}=1, R^2 (POC, EC) =0.67±0.03). The scatter plots demonstrate regression examples from a single run. The box plots show the distribution of regressed slopes and intercepts from 5000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

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2017-300
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 27 September 2017



599

600 **Figure 6.** Regression results using ambient σ_{abs520} and EC data from a suburban site in Guangzhou, China.

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2017-300
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 27 September 2017

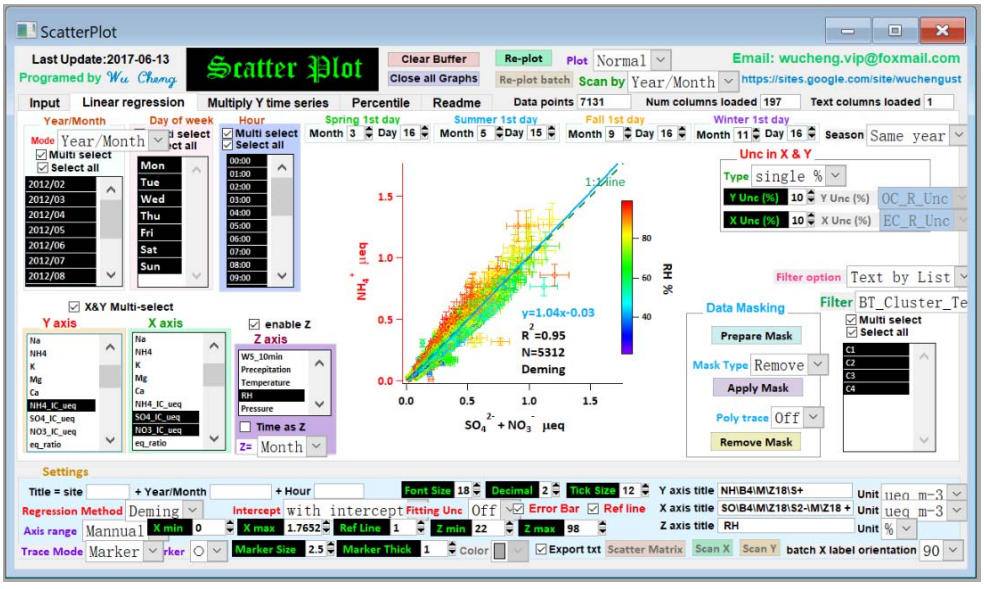


Figure 7. The user interface of Scatter Plot Igor program. The program and its operation manual are available from: https://sites.goo le.com/site/wuchengust.