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)

17 **Abstract**

18 Linear regression techniques are widely used in atmospheric science, but are often
19 improperly applied due to lack of consideration or inappropriate handling of
20 measurement uncertainty. In this work, numerical experiments are performed to
21 evaluate the performance of five linear regression techniques, significantly extending
22 previous works by Chu and Saylor. The five techniques are Ordinary Least Square
23 (OLS), Deming Regression (DR), Orthogonal Distance Regression (ODR), Weighted
24 ODR (WODR), and York regression (YR). We first introduce a new data generation
25 scheme that employs the Mersenne Twister (MT) pseudorandom number generator.
26 The numerical simulations are also improved by: (a) refining the parameterization of
27 non-linear measurement uncertainties, (b) inclusion of a linear measurement
28 uncertainty, (c) inclusion of WODR for comparison. Results show that DR, WODR and
29 YR produce an accurate slope, but the intercept by WODR and YR is overestimated
30 and the degree of bias is more pronounced with a low R^2 XY dataset. The importance
31 of a properly weighting parameter λ in DR is investigated by sensitivity tests, and it is
32 found that an improper λ in DR can lead to a bias in both the slope and intercept
33 estimation. Because the λ calculation depends on the actual form of the measurement
34 error, it is essential to determine the exact form of measurement error in the XY data
35 during the measurement stage. If a priori error in one of the variables is unknown, or
36 the measurement error described cannot be trusted, DR, WODR and YR can provide
37 the least biases in slope and intercept among all tested regression techniques. For these
38 reasons, DR, WODR and YR are recommended for atmospheric studies when both X
39 and Y data have measurement errors.
40

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_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; 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; Chow et al., 2004; Cheng et al., 2011; Wu et al., 2016), light extinction budget reconstruction (Malm et al., 1994; Watson, 2002; Li et al., 2017), comparison between modeling and measurement (Petäjä et al., 2009), emission factor study (Janhäll et al., 2010), retrieval of shortwave cloud forcing (Cess et al., 1995), calculation of pollutant growth rate (Richter et al., 2005), estimation of ground level $PM_{2.5}$ from MODIS data (Wang and Christopher, 2003), distinguishing OC origin from biomass burning using K^+ as a tracer (Duan et al., 2004) and emission type identification by the EC/CO ratio (Chen et al., 2001).

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

To overcome the drawback of OLS, a number of error-in-variable regression models (also known as bivariate fittings (Cantrell, 2008) or total least-squares methods (Markovsky and Van Huffel, 2007) arise. Deming (1943) proposed an approach by

72   minimizing sum of squares of X and Y residuals. A closed-form solution of Deming

73   regression (DR) was provided by York (1966). Method comparison work of various

74   regression techniques by Cornbleet and Gochman (1979) found significant error in OLS

75   slope estimation when the relative standard deviation (RSD) of measurement error in

76   "X" exceeded 20%, while DR was found to reach a more accurate slope estimation. In

77   an early application of the EC tracer method, Turpin and Huntzicker (1995) realized

78   the limitation of OLS since OC and EC have comparable measurement uncertainty,

79   thus recommended the use of DR for $(OC/EC)_{pri}$ (primary OC to EC ratio) estimation.

80   Ayers (2001) conducted a simple numerical experiment and concluded that reduced

81   major axis regression (RMA) is more suitable for air quality data regression analysis.

82   Linnet (1999) pointed out that when applying DR for inter-method (or inter-instrument)

83   comparison, special attention should be paid to the sample size. If the range ratio

84   (max/min) is relatively small (e.g., less than 2), more samples are needed to obtain

85   statistically significant results.

86   In principle, a best-fit regression line should have greater dependence on the more

87   precise data points rather than the less reliable ones. Chu (2005) performed a

88   comparison study of OLS and DR specifically focusing on the EC tracer method

89   application, and found the slope estimated by DR is closer to the correct value than

90   OLS but may still overestimate the ideal value. Saylor et al. (2006) extended the

91   comparison work of Chu (2005) by including a regression technique developed by York

92   et al. (2004). They found that the slope overestimation by DR in the study of Chu (2005)

93   was due to improper configuration of the weighting parameter, λ. This λ value is the

94   key to handling the uneven errors between data points for the best-fit line calculation.

95   This example demonstrates the importance of appropriate weighting in the calculation

96   of best-bit line for error-in-variable regression model, which is overlooked in many

97   studies.

98   In this study, we extend the work by Saylor et al. (2006) to achieve four objectives.

99   The first is to propose a new data generation scheme by applying the Mersenne Twister

100   (MT) pseudorandom number generator for evaluation of linear regression techniques.

101   In the study of Chu (2005), data generation is achieved by a varietal sine function,

102   which has limitations in sample size, sample distribution, and nonadjustable correlation

103   (R^2) between X and Y. In comparison, the MT data generation provides more

104   flexibility, permitting adjustable sample size, XY correlation and distribution. The

105   second is to develop a non-linear measurement error parameterization scheme for use

106   in the regression method. The third is to incorporate linear measurement errors in the

107   regression methods. In the work by Chu (2005) and Saylor et al. (2006), the relative

108   measurement uncertainty (γ_{Unc}) is non-linear with concentration, but a constant γ_{Unc}

109   is often applied on atmospheric instruments due to its simplicity. The fourth is to

110   include weighted orthogonal distance regression (WODR) for comparison.

111   Abbreviations and symbols used in this study are summarized in Table 1 for quick

112   reference.

113   **2   Description of regression techniques compared in this study**

114   **Ordinary least squares (OLS) method**. OLS only considers the errors in dependent

115   variables (Y). OLS regression is achieved by minimizing the sum of squares (S) in the

116   Y residuals (i.e., distance of AB in Fig. S1):

117
$$S = \sum_{i=1}^{N}(y_i - Y_i)^2 \tag{1}$$

118   where Y_i are observed Y data points while y_i are regressed Y data points of the

119   regression line. N represents the number of data points that is used for regression.

120   **Orthogonal distance regression (ODR)**. ODR minimizes the sum of the squared

121   orthogonal distances from all data points to the regressed line and considers equal error

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

123
$$S = \sum_{i=1}^{N}[(x_i - X_i)^2 + (y_i - Y_i)^2] \tag{2}$$

124   **Weighted orthogonal distance regression (WODR)**. Unlike ODR that considers even

125   error in X and Y, weightings based on measurement errors in both X and Y are

126   considered in WODR when minimizing the sum of squared orthogonal distance from

127   the data points to the regression line (Carroll and Ruppert, 1996) as shown by AD in

128   Fig. S1:

129
$$S = \sum_{i=1}^{N}[(x_i - X_i)^2 + (y_i - Y_i)^2/\eta] \tag{3}$$

130   where η is error variance ratio that determines the angle θ shown in Fig. S1.

131   Implementation of ODR and WODR in Igor Pro (WaveMetrics, Inc. Lake Oswego, OR,

132   USA) was done by the computer routine ODRPACK95 (Boggs et al., 1989; Zwolak et

133   al., 2007).

134　**Deming regression (DR)**. Deming (1943) proposed the following function to minimize

135　both the X and Y residuals as shown by AD in Fig. S1,

136　$$S = \sum_{i=1}^{N}[\omega(X_i)(x_i - X_i)^2 + \omega(Y_i)(y_i - Y_i)^2] \tag{4}$$

137　where X_i and Y_i are observed data points and x_i and y_i are regressed data points.

138　Individual data points are weighted based on errors in X_i and Y_i,

139　$$\omega(X_i) = \frac{1}{\sigma_{X_i}^2}, \quad \omega(Y_i) = \frac{1}{\sigma_{Y_i}^2} \tag{5}$$

140　where σ_{X_i} and σ_{Y_i} are the standard deviation of the error in measurement of X_i and Y_i,

141　respectively. The closed form solutions for slope and intercept of DR are shown in

142　Appendix A.

143　**York regression (YR)**. The York method (York et al., 2004) introduces the correlation

144　coefficient of errors in X and Y into the minimization function.

145　$$S = \sum_{i=1}^{N}[\omega(X_i)(x_i - X_i)^2 - 2r_i\sqrt{\omega(X_i)\omega(Y_i)}(x_i - X_i)(y_i - Y_i) + \omega(Y_i)(y_i - $$

146　$$Y_i)^2]\frac{1}{1-r_i^2} \tag{6}$$

147　where r_i is the correlation coefficient between measurement errors in X_i and Y_i. The

148　slope and intercept of YR are calculated iteratively through the formulas in Appendix

149　A.

150　Summary of the five regression techniques is given in Table S1. It is worth noting that

151　OLS and DR have closed-form expressions for calculating slope and intercept. In

152　contrast, ODR, WODR and YR need to be solved iteratively. This need to be taken into

153　consideration when choosing regression algorithm for handling huge amount of data.

154　A computer program (Scatter plot; Wu, 2017a) with graphical user interface (GUI) in

155　Igor Pro (WaveMetrics, Inc. Lake Oswego, OR, USA) is developed to facilitate the

156　implementation of error-in-variables regression (including DR, WODR and YR). Two

157　other Igor Pro based computer programs, Histbox (Wu, 2017b) and Aethalometer data

158　processor (Wu, 2017c) are used for data analysis and visualization in this study.

159　**3　Data description**

160　Two types of data are used for regression comparison. The first type is synthetic data

161　generated by computer programs, which can be used in the EC tracer method (Turpin

6

162 and Huntzicker, 1995) to demonstrate the regression application. The true "slope" and

163 "intercept" are assigned during data generation, allowing quantitative comparison of

164 the bias of each regression scheme. The second type of data comes from ambient

165 measurement of light absorption, OC and EC in Guangzhou for demonstration in a real-

166 world application.

3.1 Synthetic XY data generation

168 In this study, numerical simulations are conducted in Igor Pro (WaveMetrics, Inc. Lake

169 Oswego, OR, USA) through custom codes. Two types of generation schemes are

170 employed, one is based on the MT pseudorandom number generator (Matsumoto and

171 Nishimura, 1998) and the other is based on the sine function described by Chu (2005).

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

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

174 Here k is the regressed slope and b is the intercept. The underlying meaning is that, Y

175 can be decomposed into two parts. One part is correlated with X, and the ratio is defined

176 by k. The other part of Y is constant and independent of X and regarded as b.

177 To make the discussion easier to follow, we intentionally avoid discussion using the

178 abstract general form and instead opt to use a real-world application case in atmospheric

179 science. Linear regression had been heavily applied on OC and EC data, here we use

180 OC and EC data as an example to demonstrate the regression application in atmospheric

181 science. In the EC tracer method, OC (mixture) is Y and EC (tracer) is X. OC can be

182 decomposed into three components based on their formation pathway:

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

184 Here POC_{comb} is primary OC from combustion. $POC_{non-comb}$ is primary OC emitted from

185 non-combustion activities. SOC is secondary OC formed during atmospheric aging.

186 Since POC_{comb} is co-emitted with EC and well correlated with each other, their

187 relationship can be parameterized as:

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

189 By carefully selecting an OC and EC subset when SOC is very low (considered as

190 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 $POC_{non-comb}$ (b in Eq. (7)). With known $(OC/EC)_{pri}$ and $POC_{non-comb}$, SOC can be estimated by:

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

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

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

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

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

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

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

220 $$\gamma_{Unc-nonlinear} = \frac{1}{\sqrt{Conc._{true}}}$$ (14)

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

222 $$-\frac{1}{\sqrt{POC_{true}}} \times POC_{true} \leq \varepsilon_{POC} \leq +\frac{1}{\sqrt{POC_{true}}} \times POC_{true}$$ (15)

223 $$-\frac{1}{\sqrt{EC_{true}}} \times EC_{true} \leq \varepsilon_{EC} \leq +\frac{1}{\sqrt{EC_{true}}} \times EC_{true}$$ (16)

224 In Eq. (14), the γ_{Unc} decreases as concentration increases, since low concentrations are
225 usually more challenging to measure. As a result, the $\gamma_{Unc-nonlinear}$ defined in Eq.
226 (14) is more realistic than the constant approach, but there are two limitations. First, the
227 physical meaning of the uncertainty unit is lost. If the unit of OC is µg m^{-3}, then the
228 unit of ε_{OC} becomes $\sqrt{\mu g\ m^{-3}}$. Second, the concentration is not normalized by a
229 consistent relative value, making it sensitive to the X and Y units used. For example, if
230 POC$_{true}$=0.9 µg m^{-3}, then $\varepsilon_{POC}= \pm0.95$ µg m^{-3} and $\gamma_{Unc} = 105\%$, but by changing the
231 concentration unit to POC$_{true}$=900 ng m^{-3}, then $\varepsilon_{OC}= \pm30$ ng m^{-3} and $\gamma_{Unc} = 3\%$. To
232 overcome these deficiencies, we propose to modify Eq. (14) to:

233 $$\gamma_{Unc} = \sqrt{\frac{LOD}{Conc._{true}}} \times \alpha$$ (17)

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

241 $$-\sqrt{\frac{LOD_{POC}}{POC_{true}}} \times \alpha_{POC} \times POC_{true} \leq \varepsilon_{POC} \leq +\sqrt{\frac{LOD_{POC}}{POC_{true}}} \times \alpha_{POC} \times POC_{true}$$

242 (18)

243 $$-\sqrt{\frac{LOD_{EC}}{EC_{true}}} \times \alpha_{EC} \times EC_{true} \leq \varepsilon_{EC} \leq +\sqrt{\frac{LOD_{EC}}{EC_{true}}} \times \alpha_{EC} \times EC_{true}$$ (19)

244 With the modified $\gamma_{Unc-nonlinear}$ parameterization, concentrations of POC and EC are
245 normalized by a corresponding LOD, which maintains unit consistency between

246   POC_true and ε_{POC} and EC_true and ε_{EC}, and eliminates dependency on the concentration

247   unit.

248   Uniform distribution has been used in previous studies (Cox et al., 2003; Chu, 2005;

249   Saylor et al., 2006) and is adopted in this study to parameterize measurement error. For

250   a uniform distribution in the interval [a,b], the variance is $\frac{1}{12}(a-b)^2$. Since ε_{POC} and

251   ε_{EC} follow a uniform distribution in the interval as given by Eqs. (18) and (19), the

252   weights in DR and YR (inverse of variance) become:

253
$$\omega(X_i) = \frac{1}{\sigma_{X_i}^2} = \frac{3}{EC_{true} \times LOD_{EC} \times \alpha_{EC}^2} \tag{20}$$

254
$$\omega(Y_i) = \frac{1}{\sigma_{Y_i}^2} = \frac{3}{POC_{true} \times LOD_{POC} \times \alpha_{POC}^2} \tag{21}$$

255   The parameter λ in Deming regression is then determined:

256
$$\lambda = \frac{\omega(X_i)}{\omega(Y_i)} = \frac{POC_{true} \times LOD_{POC} \times \alpha_{POC}^2}{EC_{true} \times LOD_{EC} \times \alpha_{EC}^2} \tag{22}$$

257   Besides the $\gamma_{Unc-nonlinear}$ discussed above, a second type measurement uncertainty

258   parameterized by a constant proportional factor, $\gamma_{Unc-linear}$, is very common in

259   atmospheric applications:

260
$$-\gamma_{POCunc} \times POC_{true} \leq \varepsilon_{POC} \leq +\gamma_{POCunc} \times POC_{true} \tag{23}$$

261
$$-\gamma_{ECunc} \times EC_{true} \leq \varepsilon_{EC} \leq +\gamma_{ECunc} \times EC_{true} \tag{24}$$

262   where γ_{POCunc} and γ_{ECunc} are the relative measurement uncertainties, e.g., for relative

263   measurement uncertainty of 10%, γ_{Unc}=0.1. As a result, the measurement error is

264   linearly proportional to the concentration. An example comparison of $\gamma_{Unc-nonlinear}$

265   and $\gamma_{Unc-linear}$ is shown in Fig. 1b. For $\gamma_{Unc-linear}$, the weights become:

266
$$\omega(X_i) = \frac{1}{\sigma_{X_i}^2} = \frac{3}{(\gamma_{ECunc} \times EC_{true})^2} \tag{25}$$

267
$$\omega(Y_i) = \frac{1}{\sigma_{Y_i}^2} = \frac{3}{(\gamma_{POCunc} \times POC_{true})^2} \tag{26}$$

268   and λ for Deming regression can be determined:

269
$$\lambda = \frac{\omega(X_i)}{\omega(Y_i)} = \frac{(\gamma_{POCunc} \times POC_{true})^2}{(\gamma_{ECunc} \times EC_{true})^2} \tag{27}$$

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 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 ε_{POC} and ε_{EC}. As discussed in sect. 3.1.1, two types of measurement errors are considered for ε_{POC} and ε_{EC} calculation: $\gamma_{Unc-nonlinear}$ and $\gamma_{Unc-linear}$. In the

302 last step, POC_{measured} and EC_{measured} are calculated following Eq. (12), i.e., applying

303 measurement errors on POC_{true} and EC_{true}. Then POC_{measured} and EC_{measured} can be used

304 as Y and X, respectively, to test the performance of various regression techniques. An

305 Igor Pro based program with graphical user interface (GUI) is developed to facilitate

306 the MT data generation for OC and EC. A brief introduction is given in the

307 Supplemental Information.

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

309 Beside MT, inclusion of the sine function data generation scheme in this study mainly

310 serves two purposes. First, the sine function scheme was adopted in two previous

311 studies (Chu, 2005; Saylor et al., 2006), the inclusion of this scheme can help to verify

312 whether the codes in Igor for various regression approaches yield the same results from

313 the two previous studies. Second, the crosscheck between results from sine function

314 and MT provides circumstantial evidence that the MT scheme works as expected.

315 In this section, XY data generation by sine functions is demonstrated using POC as Y

316 and EC as X. There are four steps in POC and EC data generation as shown by the

317 flowchart in Fig. S2. Details are explained as follows: (1) The first step is to generate

318 POC and EC (Chu, 2005):

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

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

321 Here x is the elapsed hour (x=1,2,3……n; n≤120), τ is used to adjust the width of each

322 peak, and ϕ is used to adjust the phase of the sine wave. The constants 14 and 3.5 are

323 used to lift the sine wave to the positive range of the Y axis. An example of data

324 generation by the sine functions of Chu (2005) is shown in Fig. 3. Dividing Eq. (28) by

325 Eq. (29) yields a value of 4. In this way the exact relation between POC and EC is

326 defined clearly as (OC/EC)_{pri} = 4. (2) With POC_{comb} and EC_{true} generated, the second

327 step is to add POC_{non-comb} to POC_{comb} to compute POC_{true}. As for POC_{non-comb}, a single

328 value is assigned and added to all POC following Eq. (10). Then the goodness of the

329 regression intercept can be evaluated by comparing the regressed intercept with preset

330 POC_{non-comb}. (3) The third step is to compute ε_{POC} and ε_{EC} , considering both

331 $\gamma_{Unc-nonlinear}$ and $\gamma_{Unc-linear}$. (4) The last step is to apply measurement errors on

332   POC$_{true}$ and EC$_{true}$ following Eq. (12). Then POC$_{measured}$ and EC$_{measured}$ can be used as

333   Y and X, respectively, to evaluate the performance of various regression techniques.

3.2   Ambient measurement of σ$_{abs}$ and EC

335   Sampling was conducted from Feb 2012 to Jan 2013 at the suburban Nancun (NC) site

336   (23° 0'11.82"N, 113°21'18.04"E), which is situated on top of the highest peak (141 m

337   ASL) in the Panyu district of Guangzhou. This site is located at the geographic center

338   of Pearl River Delta region (PRD), making it a good location for representing the

339   average atmospheric mixing characteristics of city clusters in the PRD region. Light

340   absorption measurements were performed by a 7λ Aethalometer (AE-31, Magee

341   Scientific Company, Berkeley, CA, USA). EC mass concentrations were measured by

342   a real time ECOC analyzer (Model RT-4, Sunset Laboratory Inc., Tigard, Oregon,

343   USA). Both instruments utilized inlets with a 2.5 μm particle diameter cutoff. The algorithm

344   of Weingartner et al. (2003) was adopted to correct the sampling artifacts (aerosol

345   loading, filter matrix and scattering effect) (Coen et al., 2010) in Aethalometer

346   measurement. A customized computer program with graphical user interface,

347   Aethalometer data processor (Wu et al., 2018), was developed to perform the data

348   correction   and   detailed   descriptions   can   be   found   in

349   https://sites.google.com/site/wuchengust. More details of the measurements can be

350   found in Wu et al. (2018).

4   Comparison study using synthetic data

352   In the following comparisons, six regression approaches are compared using two data

353   generation schemes (Chu sine function and MT) separately, as illustrated in Fig. 4. Each

354   data generation scheme considers both $\gamma_{Unc-nonlinear}$ and $\gamma_{Unc-linear}$ in measurement

355   error parameterization. In total, 18 cases are tested with different combination of data

356   generation schemes, measurement error parameterization schemes, true slope and

357   intercept settings. In each case, six regression approaches are tested, including OLS,

358   DR ($\lambda = 1$), DR ($\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$), ODR, WODR and YR. In commercial software (e.g.,

359   OriginPro®, SigmaPlot®, GraphPad Prism®, etc), λ in DR is set to 1 by default if not

360   specified. As indicated by Saylor et al. (2006), the bias observed in the study of Chu

361   (2005) is likely due to $\lambda = 1$ in DR. The purpose of including DR ($\lambda = 1$) in this study

362 is to examine the potential bias using the default input in many software products. The

363 six regression approaches are considered to examine the sensitivity of regression results

364 to various parameters used in data generation. For each case, 5000 runs are performed

365 to obtain statistically significant results, as recommended by Saylor et al. (2006). The

366 mean slope and intercept from 5000 runs is compared with the true value assigned

367 during data generation. If the difference is <5%, the result is considered unbiased.

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

369 In this section, the scheme of Chu (2005) is adopted for data generation to obtain a

370 benchmark of six regression approaches. With different setup of slope, intercept and

371 γ_{Unc} , 6 cases (Case 1 ~ 6) are studied and the results are discussed below.

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

373 A comparison of the regression techniques results with $\gamma_{Unc-nonlinear}$ (following Eqs.

374 (18) & (19)) is summarized in Table 2. LOD_{POC} , LOD_{EC}, α_{POC} and α_{EC} are all set to 1

375 to reproduce the data studied by Chu (2005) and Saylor et al. (2006). Two sets of true

376 slope and intercept are considered (Case 1: Slope=4, Intercept=0; Case 2: Slope=4,

377 Intercept=3) to examine if any results are sensitive to the non-zero intercept. The R^2

378 (POC, EC) from 5000 runs for both case 1 and 2 are 0.67±0.03.

379 As shown in Fig. 5, for the zero-intercept case (Case 1), OLS significantly

380 underestimates the slope (2.95±0.14) while overestimates the intercept (5.84±0.78).

381 This result indicates that OLS is not suitable for errors-in-variables linear regression,

382 consistent with similar analysis results from Chu (2005) and Saylor et al. (2006). With

383 DR, if the λ is properly calculated by weights ($\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$), unbiased slope (4.01±0.25)

384 and intercept (-0.04±1.28) are obtained; however, results from DR with λ=1 show

385 obvious bias in the slope (4.27±0.27) and intercept (-1.45±1.36). ODR also produces

386 biased slope (4.27±0.27) and intercept (-1.45±1.36), which are identical to results of

387 DR when λ=1. With WODR, unbiased slope (3.98±0.22) is observed, but the intercept

388 is overestimated (1.12±1.02). Results of YR are identical to WODR. For Case 2

389 (slope=4, intercept=3), slopes from all six regression approaches are consistent with

14

390   Case 1 (Table 2). The Case 2 intercepts are equal to the Case 1 intercepts plus 3,

391   implying that all the regression methods are not sensitive to a non-zero intercept.

392   For case 3, LOD_{POC}=0.5, LOD_{EC}=0.5, α_{POC}=0.5, α_{EC}=0.5 are adopted (Table 2),

393   leading to an offset to the left of $\gamma_{Unc-nonlinear}$ (blue curve) compared to Case 1 and 2

394   (black curve) in Fig. 1. As a result, for the same concentration of EC and OC in Case

395   3, the $\gamma_{Unc-nonlinear}$ is smaller than in Case 1 and Case 2 as indicated by a higher R^2

396   (0.95±0.01 for Case 3, Table 2). With a smaller measurement uncertainty, the degree

397   of bias in Case 3 is smaller than in Case 1. For example, OLS slope is less biased in

398   Case 3 (3.83±0.08) compared to Case 1 (2.94±0.14). Similarly, the slope (4.03±0.09)

399   and intercept (-0.18±0.44) of DR (λ=1) exhibit a much smaller bias with a smaller

400   measurement uncertainty, implying that the degree of bias by improperly weighting in

401   DR, WODR and YR is associated with the degree of measurement uncertainty. A higher

402   measurement uncertainty results in larger bias in slope and intercept.

403   An uneven LOD_{POC} and LOD_{EC} is tested in Case 4 with LOD_{POC}=1, LOD_{EC}=0.5,

404   α_{POC}=0.5, α_{EC}=0.5, which yield a R^2(POC, EC) of 0.78±0.02. The results are similar

405   to Case 1. For DR ($\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$) unbiased slope and intercept are obtained. For WODR

406   and YR, unbiased slopes are reported with a small bias in the intercepts. Large bias

407   values are observed in both the slopes and intercepts in Case 4 using OLS, DR ($\lambda = 1$)

408   and ODR.

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

410   Cases 5 and 6 represent the results from using $\gamma_{Unc-linear}$ and are shown in Table 2.

411   γ_{Unc} is set to 30% to achieve a R^2 (POC, EC) of 0.7, a value close to the R^2 in studies

412   of Chu (2005) and Saylor et al. (2006). In Case 5 (slope=4, intercept=0), unbiased

413   slopes and intercepts are determined by DR ($\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$), WODR and YR. OLS

414   underestimates the slope (3.32 ±0.20) and overestimates intercept (3.77 ±0.90), while

415   DR ($\lambda = 1$) and ODR overestimate the slopes (4.75 ±0.30) and underestimate the

416   intercepts (-4.14 ±1.36). In Case 6 (slope=4, intercept=3), results similar to Case 5 are

417   obtained. It is worth noting that although the mean intercept (3.05±1.22) of DR ($\lambda =$

418     $\frac{\omega(X_i)}{\omega(Y_i)}$), is closest to the true value (intercept=3), the deviations are much larger than for

419   WODR (2.72±0.74).

4.2   Comparison results using data generated by MT

421 In this section, MT is adopted for data generation to obtain a benchmark of six

422 regression approaches. Both $\gamma_{Unc-nonlinear}$ and $\gamma_{Unc-linear}$ are considered. With

423 different configuration of slope, intercept and γ_{Unc} , 12 cases (Case 7 ~ Case 18) are

424 studied and the results are discussed below.

4.2.1 $\gamma_{Unc-nonlinear}$ results

426 Cases 7 and 8 use data generated by MT and $\gamma_{Unc-nonlinear}$ with results shown in Table

427 2. In Case 7 (slope=4, intercept=0, LOD_{POC}=1, LOD_{EC}=1, α_{POC}=1, α_{EC}=1), unbiased

428 slope (4.00 ±0.03) and intercept (0.00 ±0.17) is estimated by DR ($\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$). WODR

429 and YR yield unbiased slopes (3.96 ±0.03) but overestimate the intercepts (1.21 ±0.13).

430 DR ($\lambda = 1$) and ODR report slightly biased slopes (4.17 ±0.04) with biased intercepts

431 (-0.94 ±0.18). OLS underestimates the slope (3.22 ±0.03) and overestimates the

432 intercept (4.30 ±0.14). In Case 8 (slope=4, intercept=3, LOD_{POC}=1, LOD_{EC}=1, α_{POC}=1,

433 α_{EC}=1), DR ($\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$) provides unbiased slope (4.00 ±0.03) and intercept (3.00 ±0.18)

434 estimations. WODR and YR report unbiased slopes (3.97 ±0.03) and overestimate

435 intercepts (4.11 ±0.13). OLS, DR ($\lambda = 1$) and ODR report biased slopes and intercepts.

436 To test the overestimation/underestimation dependency on the true slope, Case 9

437 (slope=0.5, intercept=0, LOD_{POC} =1, LOD_{EC} =1, α_{POC} =1, α_{EC} =1) and case 10

438 (slope=0.5, intercept=3, LOD_{POC}=1, LOD_{EC}=1, α_{POC}=1, α_{EC}=1) are conducted and the

439 results are shown in Table 2. Unlike the overestimation observed in Case 1~Case 8, DR

440 ($\lambda = 1$) and ODR underestimate the slopes (0.46 ±0.01) in Case 9. In case 10, DR ($\lambda = $

441 1), DR ($\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$) and ODR report unbiased slopes and intercepts. Case 11 and case

442 12 test the bias when the true slope is 1 as shown in Table 2. In Case 11 (intercept=0),

443 all regression approaches except OLS can provide unbiased results. In Case 12, all

444 regression approaches report unbiased slopes except OLS, but DR ($\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$) is the

445 only regression approach that reports unbiased intercept.

446 These results imply that if the true slope is less than 1, the improper weighting ($\lambda = 1$)

447 in Deming regression and ODR without weighting tends to underestimate slope. If the

448 true slope is 1, these two estimators can provide unbiased results. If the true slope is

449 larger than 1, the improper weighting ($\lambda = 1$) in Deming regression and ODR without

450 weighting tends to overestimate slope.

451 4.2.2 $\gamma_{Unc-linear}$ results

452 Cases 13 and 14 (Table 2) represent the results from using $\gamma_{Unc-linear}$ (30%) and data

453 generated from MT. For case 13 (slope=4, intercept=0), DR ($\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$), WODR and

454 YR provide the best estimation of slopes and intercepts. DR ($\lambda = 1$) and ODR

455 overestimate slopes (4.53 ±0.05) and underestimate intercepts (-2.94 ±0.24). For case

456 14 (slope=4, intercept=3), DR ($\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$), WODR and YR provide an unbiased

457 estimation of slopes. But DR ($\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$) is the only regression approach reporting

458 unbiased intercept (3.08 ±0.23). Cases 15 and 16 are tested to investigate whether the

459 results are different if the true slope is smaller than 1. As shown in Table 2, the results

460 are similar to case 13&14 that DR ($\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$) can provide unbiased slope and intercept

461 while WODR and YR can provide unbiased slopes but biased intercepts. Cases 17 and

462 18 are tested to see if the results are the same for a special case when the true slope is

463 1. As shown in Table 2, the results are similar to case 13&14, implying that these results

464 are not sensitive to the special case when the true slope is 1.

465 4.3 The importance of appropriate λ input for Deming regression

466 As discussed above, inappropriate λ assignment in the Deming regression (e.g., λ=1 by

467 default for much commercial software) leads to biased slope and intercept. Beside λ=1,

468 inappropriate λ input due to improper handling of measurement uncertainty can also

469 result in bias for Deming regression. An example is shown in Fig. S3. Data is generated

470 by MT with following parameters: slope=4, intercept=0, and $\gamma_{Unc-linear}$ (30%). Fig.

471 S2 a&b demonstrates that when an appropriate λ is provided (following $\gamma_{Unc-linear}$,

472 $\lambda = \frac{POC^2}{EC^2}$), unbiased slopes and intercepts are obtained. If an improper λ is used due to

473 a mismatched measurement uncertainty assumption ($\gamma_{Unc-nonlinear}$, $\lambda = \frac{POC}{EC}$), the

474   slopes are overestimated (Fig. S3c, 4.37±0.05) and intercepts are underestimated (Fig.

475   S3d，-2.01±0.24). This result emphasizes the importance of determining the correct

476   form of measurement uncertainty in ambient samples, since λ is a crucial parameter in

477   Deming regression.

478   In the λ calculation, different representations for POC and EC, including mean, median

479   and mode, are tested as shown in Fig. S4. The results show that when X and Y have a

480   similar distribution (e.g., both are log-normal), any of mean, median or mode can be

481   used for the λ calculation.

482   **4.4   Caveats of regressions with unknown X and Y uncertainties**

483   In atmospheric applications, there are scenarios in which a priori error in one of the

484   variables is unknown, or the measurement error described cannot be trusted. For

485   example, in the case of comparing model prediction and measurement data, the

486   uncertainty of model prediction data is unknown. A second example is the case in which

487   measurement uncertainty cannot be determined due to the lack of duplicated or

488   collocated measurements and as a result, an arbitrarily assumed uncertainty is used.

489   Such a case was illustrated in the study by Flanagan et al. (2006). They found that in the

490   Speciation Trends Network (STN), the whole-system uncertainty retrieved by data from

491   collocated samplers was different from the arbitrarily assumed 5% uncertainty.

492   Additionally, the discrepancy between the actual uncertainty obtained through

493   collocated samplers and the arbitrarily assumed uncertainty varied by chemical species.

494   To investigate the performance of different regression approaches in these cases, two

495   tests (A and B) are conducted.

496   In Test A, the actual measurement error for X is fixed at 30% while γ_{Unc} for Y varies

497   from 1% to 50%. The assumed measurement error for regression is 10% for both X and

498   Y. Result of Test A are shown in Figs. 6 a and b. For OLS, the slopes are under-

499   estimated (-14 ~ -12%) and intercepts are overestimated (90 ~ 103%) and the biases are

500   independent of variations in γ_{Unc_Y}. ODR and DR ($\lambda = 1$) yield similar results with

501   over-estimated slopes (0 ~ 44%) and under-estimated intercepts (-330 ~ 0%). The degree

502   of bias in slopes and intercepts depends on the γ_{Unc_Y}. WODR, DR ($\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$) and YR

18

503 perform much better than other regression approaches in Test A, with a smaller bias in
504 both slopes (-8 ~ 12%) and intercepts -98 ~ 55%).

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

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

524 **5 Regression applications to ambient data**

525 This section demonstrates the application of the 6 regression approaches on a light
526 absorption coefficient and EC dataset collected in a suburban site in Guangzhou. As
527 mentioned in sect. 4.4, measurement uncertainties are crucial inputs for DR, YR and
528 WODR. The measurement precision of Aethalometer is 5% (Hansen, 2005) while EC
529 by RT-ECOC analyzer is 24% (Bauer et al., 2009). These measurement uncertainties
530 are used in DR, YR and WODR calculation. The data-set contains 6926 data points
531 with a R^2 of 0.92.

532 As shown in Fig. 7, Y axis is light absorption at 520 nm (σ_{abs520}) and the X axis is EC

533 mass concentration. The regressed slopes represent the mass absorption efficiency

534 (MAE) of EC at 520 nm, ranging from 13.66 to 15.94 m^2g^{-1} by the six regression

535 approaches. OLS yields the lowest slope (13.66 as shown in Fig. 7a) among all six

536 regression approaches, consistent with the results using synthetic data. This implies that

537 OLS tends to underestimate regression slope when mean Y to X ratio is larger than 1.

538 DR ($\lambda = 1$) and ODR report the same slope (14.88) and intercept (5.54), this

539 equivalency is also observed for the synthetic data. Similarly, WODR and YR yield

540 identical slope (14.88) and intercept (5.54), in line with the synthetic data results. The

541 regressed slope by DR ($\lambda = 1$) is higher than DR ($\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$), and this relationship

542 agrees well with the synthetic data results.

543 Regression comparison is also performed on hourly OC and EC data. Regression on

544 OC/EC percentile subset is a widely used empirical approach for primary OC/EC ratio

545 determination. Fig. S5 shows the regression slopes as a function of OC/EC percentile.

546 OC/EC percentile ranges from 0.5% to 100%, with an interval of 0.5%. As the

547 percentile increases, SOC contribution in OC increases as well, resulting in decreased

548 R^2 between OC and EC. The deviations between six regression approaches exhibit a

549 dependency on R^2. When percentile is relatively small (e.g., <10%), the differences

550 between the six regression approaches are also small due to the high R^2 (0.98). The

551 deviations between the six regression approaches become more pronounced as R^2

552 decreases (e.g., <0.9). The deviations are expected to be even larger when R^2 is less

553 than 0.8. These results emphasize the importance of applying error-in-variables

554 regression, since ambient XY data more likely has a R^2 less than 0.9 in most cases.

555 As discussed in this section, the ambient data confirm the results obtained in comparing

556 methods with the synthetic data. The advantage of using the synthetic data for

557 regression approaches evaluation is that the ideal slope and intercept are known values

558 during the data generation, so the bias of each regression approach can be quantified.

559 **6   Recommendations and conclusions**

560 This study aims to provide a benchmark of commonly used linear regression algorithms

561 using a new data generation scheme (MT). Six regression approaches are tested,

562 including OLS, DR ($\lambda = 1$), DR ($\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$), ODR, WODR and YR. The results show

563 that OLS fails to estimate the correct slope and intercept when both X and Y have

564 measurement errors. This result is consistent with previous studies. For ambient data

565 with R^2 less than 0.9, error-in-variables regression is needed to minimize the biases in

566 slope and intercept. If measurement uncertainties in X and Y are determined during the

567 measurement, measurement uncertainties should be used for regression. With

568 appropriate weighting, DR, WODR and YR can provide the best results among all

569 tested regression techniques. Sensitivity tests also reveal the importance of the

570 weighting parameter λ in DR. An improper λ could lead to biased slope and intercept.

571 Since the λ estimation depends on the form of the measurement errors, it is important

572 to determine the measurement errors during the experimentation stage rather than

573 making assumptions. If measurement errors are not available from the measurement

574 and assumptions are made on measurement errors, DR, WODR and YR are still the

575 best option that can provide the least bias in slope and intercept among all tested

576 regression techniques. For these reasons, DR, WODR and YR are recommended for

577 atmospheric studies when both X and Y data have measurement errors.

578 Application of error-in-variables regression is often overlooked in atmospheric studies,

579 partly due to the lack of a specified tool for the regression implementation. To facilitate

580 the implementation of error-in-variables regression (including DR, WODR and YR), a

581 computer program (Scatter plot) with graphical user interface (GUI) in Igor Pro

582 (WaveMetrics, Inc. Lake Oswego, OR, USA) is developed (Fig. 8). It is packed with

583 many useful features for data analysis and plotting, including batch plotting, data

584 masking via GUI, color coding in Z axis, data filtering and grouping by numerical

585 values and strings. The Scatter plot program and user manual are available from

586 https://sites.google.com/site/wuchengust and https://doi.org/10.5281/zenodo.832417.

587

588 **Appendix A: Equations of regression techniques**

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

590 First calculate average of observed X_i and Y_i.

591
$$\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N} \tag{A1}$$

592
$$\bar{Y} = \frac{\sum_{i=1}^{N} Y_i}{N} \tag{A2}$$

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

594
$$S_{xx} = \sum_{i=1}^{N}(X_i - \bar{X})^2 \tag{A3}$$

595
$$S_{yy} = \sum_{i=1}^{N}(Y_i - \bar{Y})^2 \tag{A4}$$

596 OLS slope and intercept can be obtained from,

597
$$k = \frac{S_{yy}}{S_{xx}} \tag{A6}$$

598
$$b = \bar{Y} - k\bar{X} \tag{A7}$$

599

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

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

602
$$S_{xy} = \sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y}) \tag{A8}$$

603 DR slope and intercept can be obtained from,

604
$$k = \frac{S_{yy} - \lambda S_{xx} + \sqrt{\left(S_{yy} - \lambda S_{xx}\right)^2 + 4\lambda S_{xy}^2}}{2S_{xy}} \tag{A9}$$

605
$$b = \bar{Y} - k\bar{X} \tag{A10}$$

606

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

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

609
$$W_i = \frac{\omega(X_i)\omega(Y_i)}{\omega(X_i) + k^2\omega(Y_i) - 2kr_i\sqrt{\omega(X_i)\omega(Y_i)}} \tag{A11}$$

$$U_i = X_i - \bar{X} = X_i - \frac{\sum_{i=1}^{N} W_i X_i}{\sum_{i=1}^{N} W_i} \tag{A12}$$

$$V_i = Y_i - \bar{Y} = Y_i - \frac{\sum_{i=1}^{N} W_i Y_i}{\sum_{i=1}^{N} W_i} \tag{A13}$$

Then calculate β_i.

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

Slope and intercept can be obtained from,

$$k = \frac{\sum_{i=1}^{N} W_i \beta_i V_i}{\sum_{i=1}^{N} W_i \beta_i U_i} \tag{A15}$$

$$b = \bar{Y} - k\bar{X} \tag{A16}$$

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

References

636 | **References**

637 Ayers, G. P.: Comment on regression analysis of air quality data, Atmos. Environ., 35,
638 2423-2425, doi: 10.1016/S1352-2310(00)00527-6, 2001.

639 Bauer, J. J., Yu, X.-Y., Cary, R., Laulainen, N., and Berkowitz, C.: Characterization of
640 the sunset semi-continuous carbon aerosol analyzer, J. Air Waste Manage. Assoc., 59,
641 826-833, doi: 10.3155/1047-3289.59.7.826, 2009.

642 Boggs, P. T., Donaldson, J. R., and Schnabel, R. B.: Algorithm 676: ODRPACK:
643 software for weighted orthogonal distance regression, ACM Trans. Math. Softw., 15,
644 348-364, doi: 10.1145/76909.76913, 1989.

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

648 Cantrell, C. A.: Technical Note: Review of methods for linear least-squares fitting of
649 data and application to atmospheric chemistry problems, Atmos. Chem. Phys., 8, 5477-
650 5487, doi: 10.5194/acp-8-5477-2008, 2008.

651 Carroll, R. J. and Ruppert, D.: The use and misuse of orthogonal regression in linear
652 errors-in-variables models, Am. Stat., 50, 1-6, doi: 10.1080/00031305.1996.10473533,
653 1996.

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

659 Chen, L. W. A., Doddridge, B. G., Dickerson, R. R., Chow, J. C., Mueller, P. K., Quinn,
660 J., and Butler, W. A.: Seasonal variations in elemental carbon aerosol, carbon monoxide
661 and sulfur dioxide: Implications for sources, Geophys. Res. Lett., 28, 1711-1714, doi:
662 10.1029/2000GL012354, 2001.

663 Cheng, Y., Duan, F.-k., He, K.-b., Zheng, M., Du, Z.-y., Ma, Y.-l., and Tan, J.-h.:
664 Intercomparison of thermal–optical methods for the determination of organic and
665 elemental carbon: influences of aerosol composition and implications, Environ. Sci.
666 Technol., 45, 10117-10123, doi: 10.1021/es202649g, 2011.

667 Chow, J. C., Watson, J. G., Crow, D., Lowenthal, D. H., and Merrifield, T.: Comparison
668 of IMPROVE and NIOSH carbon measurements, Aerosol. Sci. Technol., 34, 23-34,
669 doi: 10.1080/027868201300081923, 2001.

670 Chow, J. C., Watson, J. G., Chen, L. W. A., Arnott, W. P., and Moosmuller, H.:
671 Equivalence of elemental carbon by thermal/optical reflectance and transmittance with
672 different temperature protocols, Environ. Sci. Technol., 38, 4414-4422, doi:
673 10.1021/Es034936u, 2004.

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

676 Coen, M. C., Weingartner, E., Apituley, A., Ceburnis, D., Fierz-Schmidhauser, R.,
677 Flentje, H., Henzing, J. S., Jennings, S. G., Moerman, M., Petzold, A., Schmid, O., and

678 Baltensperger, U.: Minimizing light absorption measurement artifacts of the
679 Aethalometer: evaluation of five correction algorithms, Atmos. Meas. Tech., 3, 457-
680 474, doi: 10.5194/amt-3-457-2010, 2010.

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

683 Cox, M., Harris, P., and Siebert, B. R.-L.: Evaluation of measurement uncertainty based
684 on the propagation of distributions using Monte Carlo simulation, Meas. Tech., 46, 824-
685 833, doi: 10.1023/B:METE.0000008439.82231.ad, 2003.

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

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

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

697 Flanagan, J. B., Jayanty, R. K. M., Rickman, J. E. E., and Peterson, M. R.: PM2.5
698 speciation trends network: evaluation of whole-system uncertainties using data from
699 sites with collocated samplers, J. Air Waste Manage. Assoc., 56, 492-499, doi:
700 10.1080/10473289.2006.10464516, 2006.

701 Hansen, A. D. A.: The Aethalometer manual, Berkeley, California,USA, Magee
702 Scientific, 2005.

703 Huang, X. H., Bian, Q., Ng, W. M., Louie, P. K., and Yu, J. Z.: Characterization of
704 PM2. 5 major components and source investigation in suburban Hong Kong: A one
705 year monitoring study, Aerosol. Air. Qual. Res., 14, 237-250, doi:
706 10.4209/aaqr.2013.01.0020, 2014.

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

710 Kuang, B. Y., Lin, P., Huang, X. H. H., and Yu, J. Z.: Sources of humic-like substances
711 in the Pearl River Delta, China: positive matrix factorization analysis of PM2.5 major
712 components and source markers, Atmos. Chem. Phys., 15, 1995-2008, doi:
713 10.5194/acp-15-1995-2015, 2015.

714 Li, Y., Huang, H. X. H., Griffith, S. M., Wu, C., Lau, A. K. H., and Yu, J. Z.:
715 Quantifying the relationship between visibility degradation and PM2.5 constituents at
716 a suburban site in Hong Kong: Differentiating contributions from hydrophilic and
717 hydrophobic organic compounds, Sci.Total.Environ., 575, 1571-1581, doi:
718 10.1016/j.scitotenv.2016.10.082, 2017.

719   Lim, L. H., Harrison, R. M., and Harrad, S.: The contribution of traffic to atmospheric
720   concentrations of polycyclic aromatic hydrocarbons, Environ. Sci. Technol., 33, 3538-
721   3542, doi: 10.1021/es990392d, 1999.

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

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

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

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

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

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

740   Petäjä, T., Mauldin, I. R. L., Kosciuch, E., McGrath, J., Nieminen, T., Paasonen, P.,
741   Boy, M., Adamov, A., Kotiaho, T., and Kulmala, M.: Sulfuric acid and OH
742   concentrations in a boreal forest site, Atmos. Chem. Phys., 9, 7435-7448, doi:
743   10.5194/acp-9-7435-2009, 2009.

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

747   Saylor, R. D., Edgerton, E. S., and Hartsell, B. E.: Linear regression techniques for use
748   in the EC tracer method of secondary organic aerosol estimation, Atmos. Environ., 40,
749   7546-7556, doi: 10.1016/j.atmosenv.2006.07.018, 2006.

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

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

755   von Bobrutzki, K., Braban, C. F., Famulari, D., Jones, S. K., Blackall, T., Smith, T. E.
756   L., Blom, M., Coe, H., Gallagher, M., Ghalaieny, M., McGillen, M. R., Percival, C. J.,
757   Whitehead, J. D., Ellis, R., Murphy, J., Mohacsi, A., Pogany, A., Junninen, H.,
758   Rantanen, S., Sutton, M. A., and Nemitz, E.: Field inter-comparison of eleven
759   atmospheric ammonia measurement techniques, Atmos. Meas. Tech., 3, 91-112, doi:
760   10.5194/amt-3-91-2010, 2010.

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

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

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

Wu, C., Ng, W. M., Huang, J., Wu, D., and Yu, J. Z.: Determination of Elemental and Organic Carbon in PM2.5 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, doi: 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, doi: 10.5194/amt-9-4547-2016, 2016.

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

Wu, C.: 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., Wu, D., and Yu, J. Z.: Quantifying black carbon light absorption enhancement with a novel statistical approach, Atmos. Chem. Phys., 18, 289-309, doi: 10.5194/acp-18-289-2018, 2018.

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

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

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, doi: 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, doi: 10.1016/j.atmosenv.2016.08.054, 2016.

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

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

811 **Table 1.** Summary of abbreviations and symbols.

| Abbreviation/symbol	Definition
α	a dimensionless adjustable factor to control the position of γ_{Unc} curve on the concentration axis
b	intercept in linear regression
β_i, U_i, V_i, W_i	intermediates in York regression calculations
γ_{Unc}	fractional measurement uncertainties relative to the true concentration (%)
DR	Deming regression
$\varepsilon_{EC}, \varepsilon_{POC}$	absolute measurement uncertainties of EC and POC
EC	elemental carbon
EC_{true}	numerically synthesized true EC concentration without measurement uncertainty
$EC_{measured}$	EC with measurement error ($EC_{true} + \varepsilon_{EC}$)
λ	$\omega(X_i)$ to $\omega(Y_i)$ ratio in Deming regression
k	slope in linear regression
LOD	limit of detection
MT	Mersenne twister pseudorandom number generator
OC	organic carbon
OC/EC	OC to EC ratio
$(OC/EC)_{pri}$	primary OC/EC ratio
$OC_{non-comb}$	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
$POC_{non-comb}$	numerically synthesized true POC from non-combustion sources (independent of EC_{true}) without considering measurement uncertainty
POC_{true}	sum of POC_{comb} and $POC_{non-comb}$ without considering measurement uncertainty
$POC_{measured}$	POC with measurement error ($POC_{true} + \varepsilon_{POC}$)
$\sigma_{X_i}, \sigma_{Y_i}$	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
τ	parameter in the sine function of Chu (2005) that adjusts the width of each peak
ϕ	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 σ_{X_i} and σ_{Y_i}, used as weights in DR calculation.

812

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

| | Data generation | | | | | Results by different regression approaches | | | | | | | | | | | |
| | | | | | | OLS | | DR λ=1 | | DR $\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$ | | ODR | | WODR | | YR | |
Case	Data scheme	True Slope	True Intercept	R² (X, Y)	Measurement error	Slope	Intercept	Slope	Intercept	Slope	Intercept	Slope	Intercept	Slope	Intercept	Slope	Intercept
1		4	0	0.67±0.03	LOD_{POC}=1, LOD_{EC}=1	2.94±0.14	5.84±0.78	4.27±0.27	-1.45±1.36	4.01±0.25	-0.04±1.28	4.27±0.27	-1.45±1.36	3.98±0.22	1.12±1.02	3.98±0.22	1.12±1.02
2		4	3	0.67±0.04	a_{POC}=1, a_{EC}=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		4	0	0.95±0.01	LOD_{POC}=0.5, LOD_{EC}=0.5 α_{POC}=0.5, α_{EC}=0.5	3.83±0.08	0.95±0.40	4.03±0.09	-0.18±0.44	4±0.09	0±0.44	4.03±0.09	-0.18±0.44	4±0.08	0.12±0.37	4±0.08	0.12±0.37
4	Chu	4	0	0.78±0.02	LOD_{POC}=1, LOD_{EC}=0.5 α_{POC}=1, α_{EC}=1	3.39±0.15	3.34±0.75	4.3±0.21	-1.66±1.06	4±0.19	-0.03±0.99	4.3±0.21	-1.66±1.06	4±0.17	0.33±0.81	4±0.17	0.33±0.81
5		4	0	0.69±0.04		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	γ_{Unc}=30%	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_{POC}=1, LOD_{EC}=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	a_{POC}=1, a_{EC}=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		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		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	γ_{Unc}=30%	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

816

817
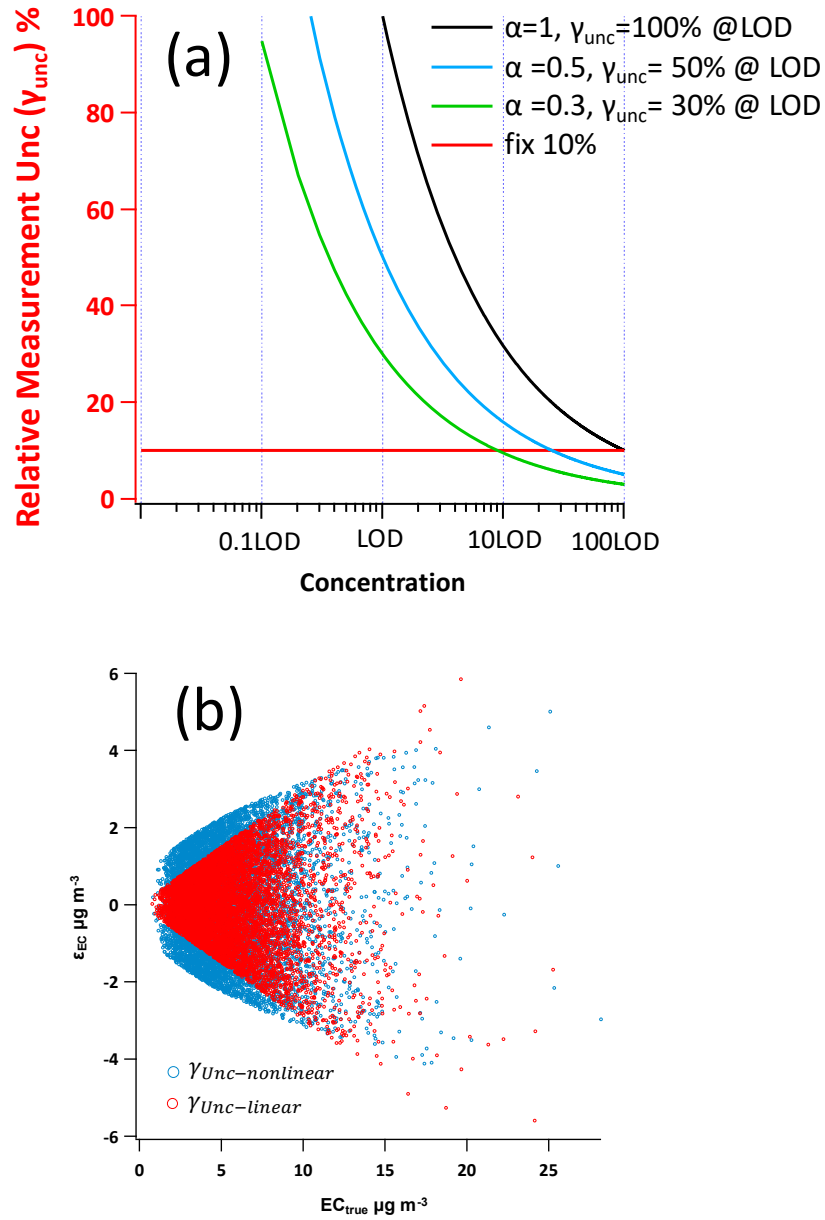
Figure 1. (a) Example $\gamma_{Unc-nonlinear}$ curves by different α values (Eq. (17)). The X
axis is concentration (normalized by LOD) in log scale and Y axis is γ_{Unc}. Black, blue
and green line represent α equal to 1, 0.5 and 0.3, respectively, corresponding to the
$\gamma_{Unc-nonlinear}$ at LOD level equals to 100%, 50% and 30%, respectively. The red line
represents $\gamma_{Unc-linear}$ of 10%. (b) Example of measurement uncertainty generation of
$\gamma_{Unc-nonlinear}$ and $\gamma_{Unc-linear}$. The blue circles represent $\gamma_{Unc-nonlinear}$ following
Eq. (17) ($LOD_{EC} = 1$, $a_{EC} = 1$). The red circles represent $\gamma_{Unc-linear}$ (30%).

825

Data generation steps by MT

827
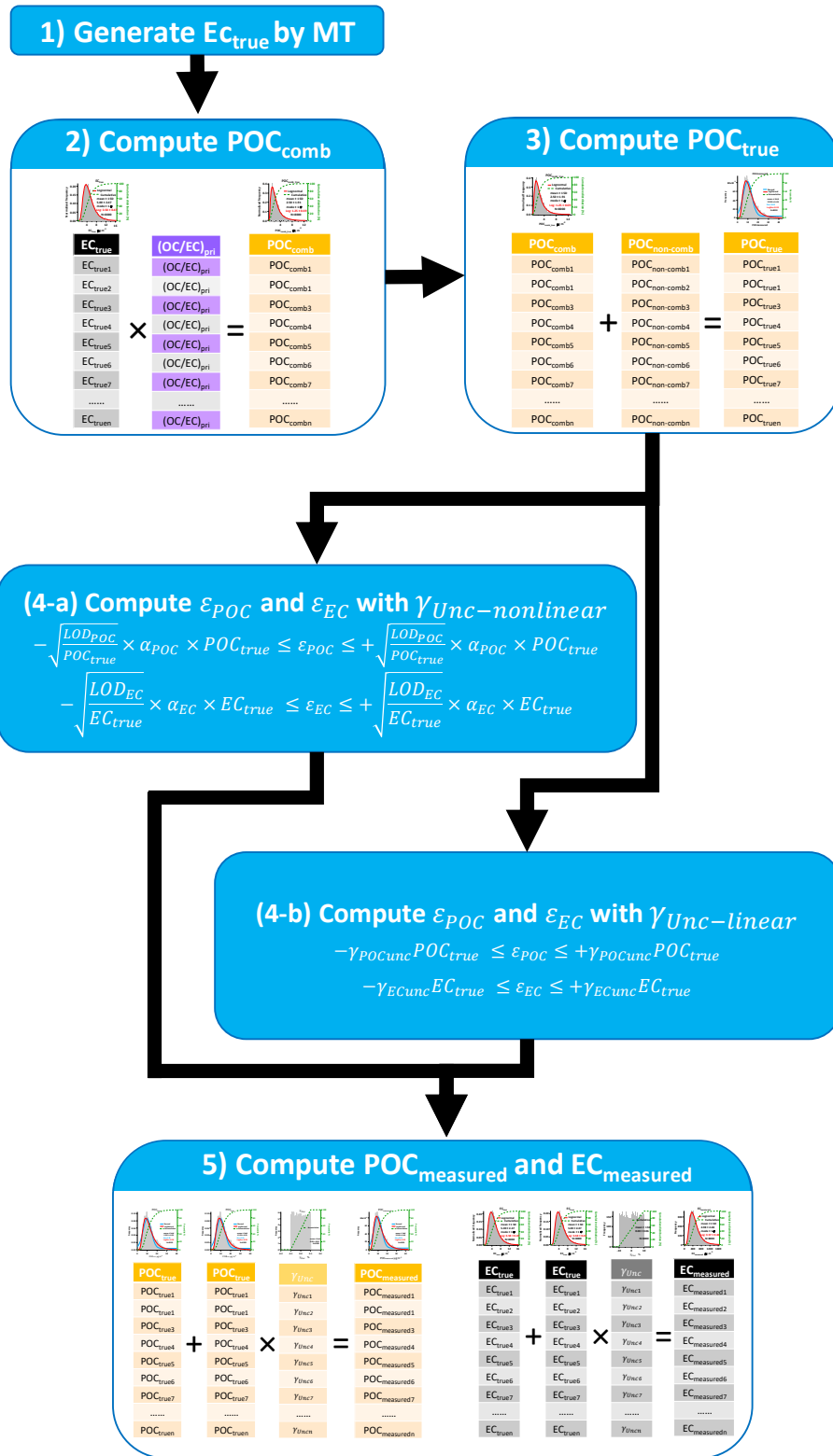
Figure 2. Flowchart of data generation steps using MT.

829
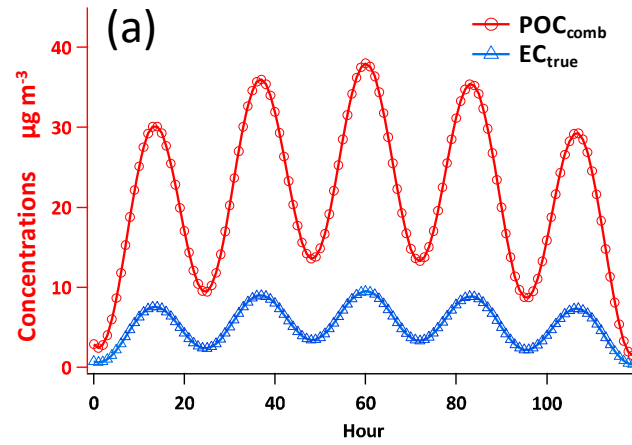
$$POC_{comb} = 14 + 12(\sin\left(\frac{x}{\tau}\right) + \sin(x - \phi))$$

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

830
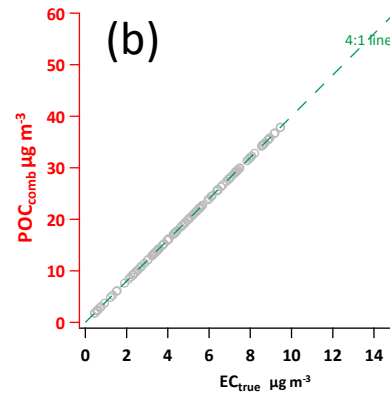
Figure 3. POC_{comb} and EC_{trure} data generated by the sine functions of Chu (2005). (a)
Time series of the 120 data points for POC_{comb} and EC_{true}. (b) Scatter plot of POC_{comb}
vs. EC_{true}

834

835

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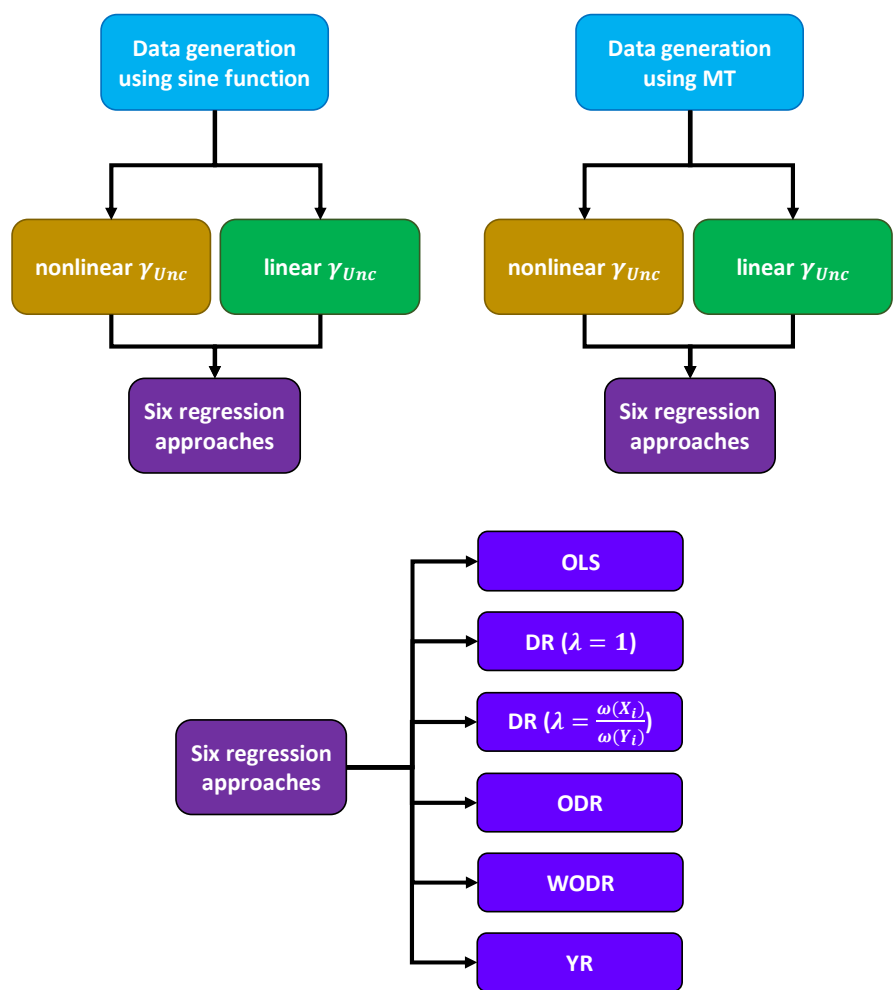
837   **Figure 4.** Overview of the comparison study design.

838
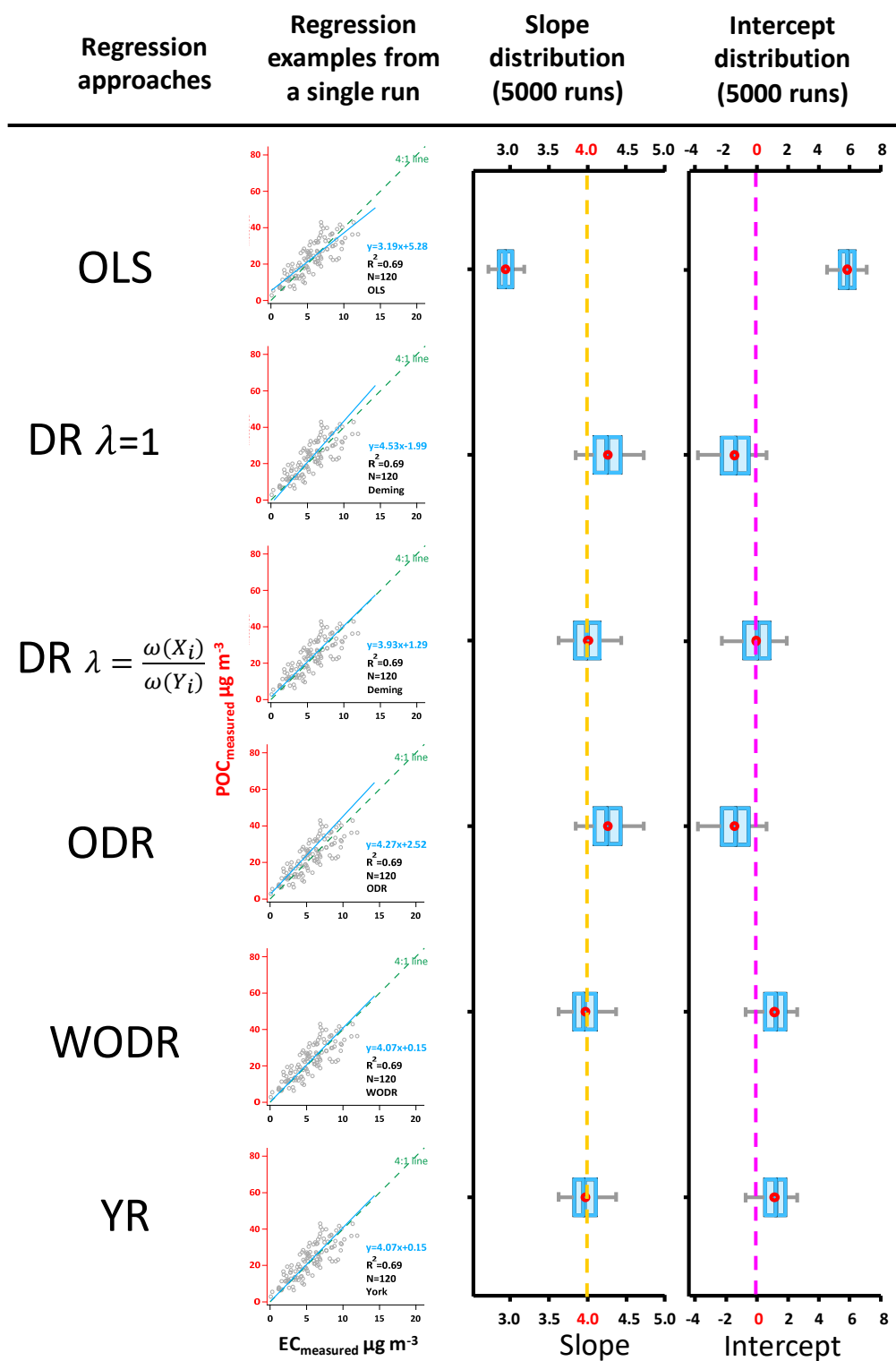
Figure 5. Regression results on synthetic data, case 1 (Slope=4, Intercept=0, LOD_{POC}=1, LOD_{EC}=1, a_{POC}=1, a_{EC}=1, R^2 (POC, EC) =0.67±0.03). The scatter plots demonstrate regression examples from a single run. The box plots show the distribution of regressed slopes and intercepts from 5000 runs of six regression approaches. The dashed line in orange and peachblow represent true slope and intercept, respectively.
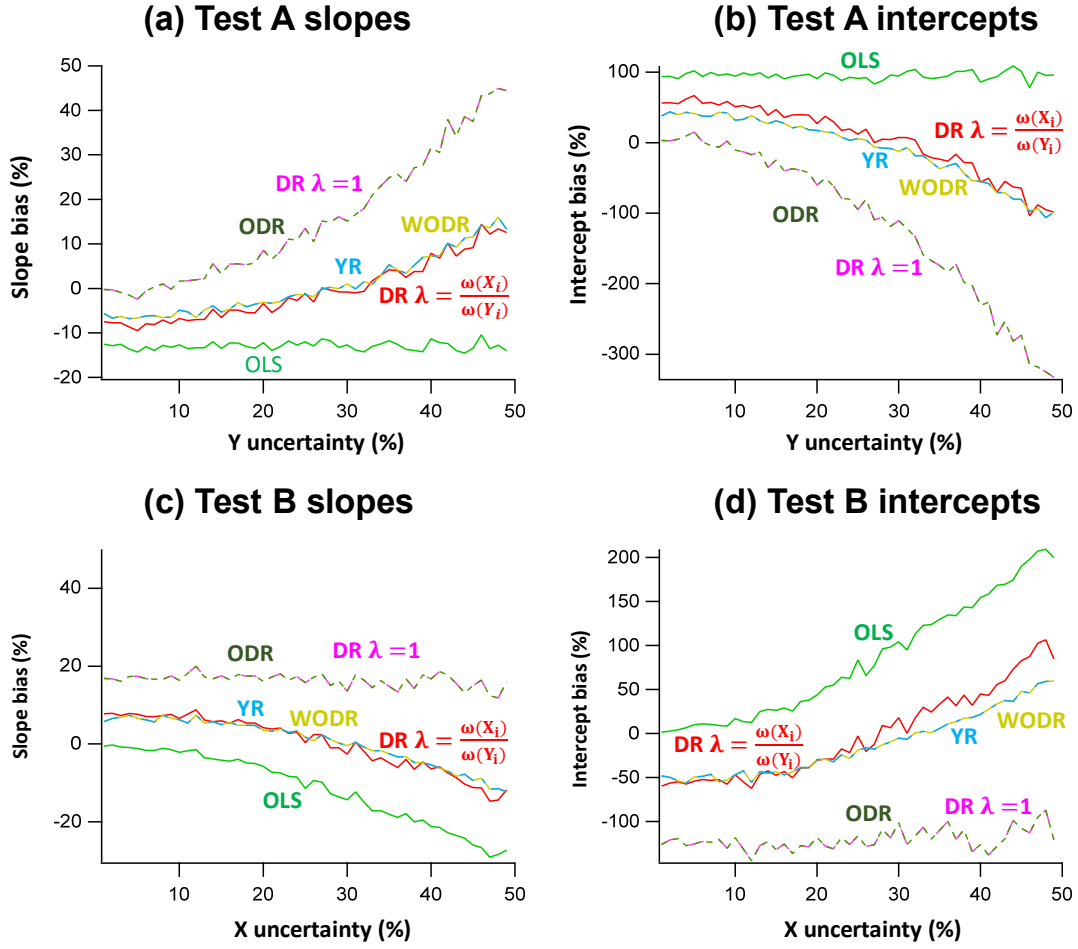
Figure 6. Slope and intercept biases by different regression schemes in two test scenarios (A and B) in which the assumed error for one of the regression variables deviates from the actual measurement error. In Test A data generation, γ_{Unc_X} is fixed at 30% and γ_{Unc_Y} is varied between 1 ~ 50%. In Test B, γ_{Unc_X} varies between 1 ~ 50% and γ_{Unc_Y} is fixed at 30%. The "true" measurement error for regression is 10% for both X and Y. (a) Slopes biases as a function of γ_{Unc_Y} in Test A. (b) Intercepts biases as a function of γ_{Unc_Y} in Test A. (c) Slopes biases as a function of γ_{Unc_X} in Test B. (d) Intercepts biases as a function of γ_{Unc_X} in Test B.
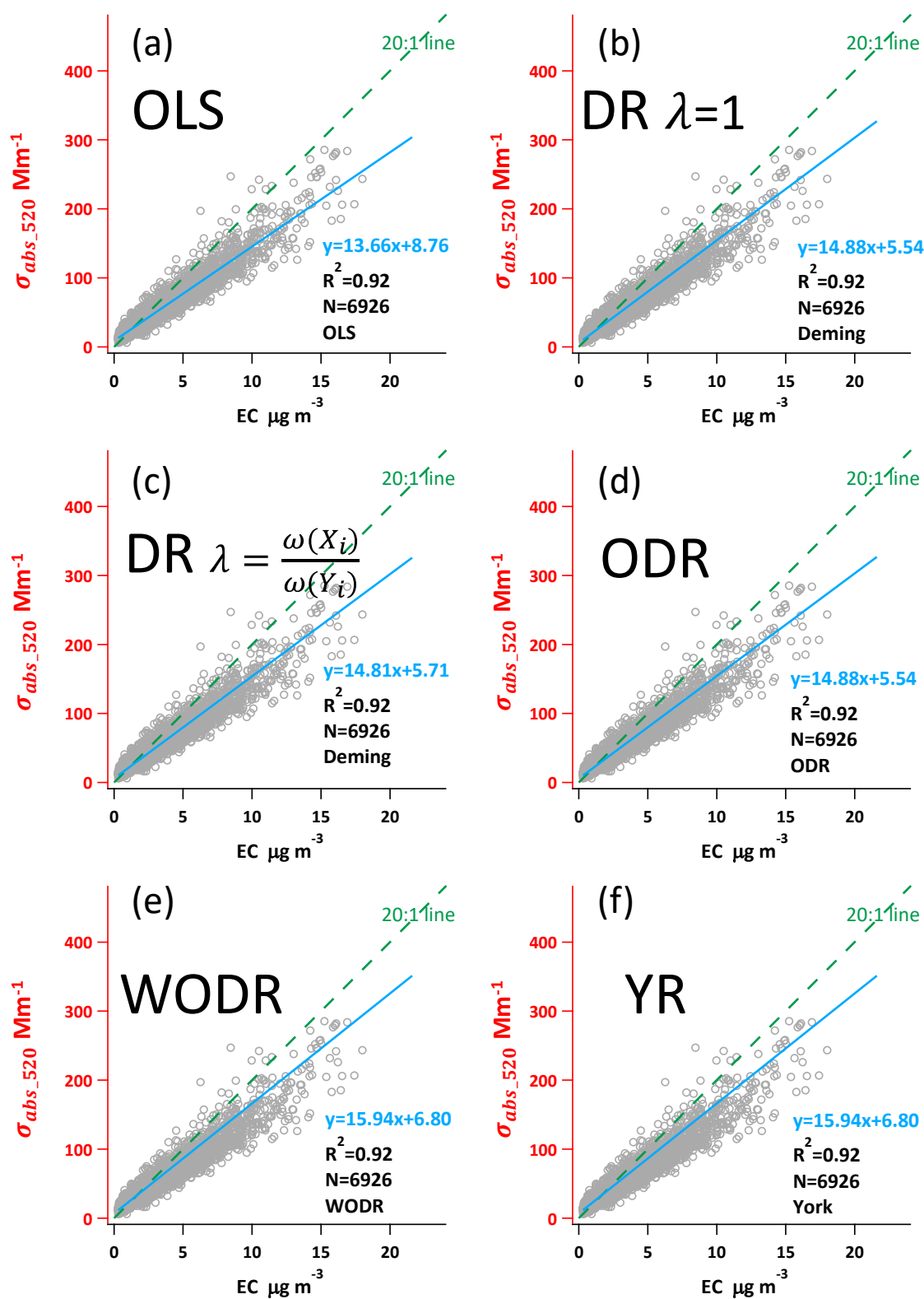
853

Figure 7. Regression results using ambient σ_{abs520} and EC data from a suburban site in Guangzhou, China.
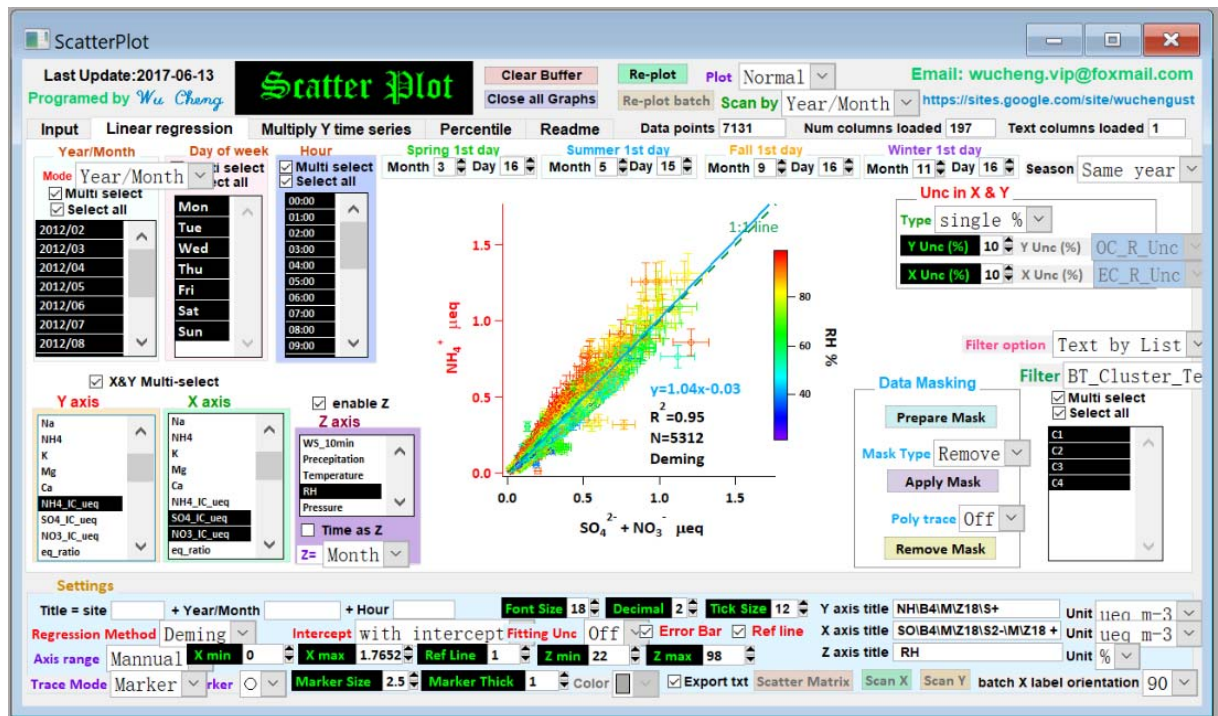
856

Figure 8. The user interface of Scatter Plot Igor program. The program and its operation

manual are available from: https://doi.org/10.5281/zenodo.832417.