Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2018-1125
Manuscript under review for journal Atmos. Chem. Phys.
Discussion started: 11 December 2018

Atmospheric Chemistry and Physics Discussions

Technical note: Effects of Uncertainties and Number of Data points on Inference from Data - a Case Study on New Particle Formation

Santtu Mikkonen[1], Mikko R. A. Pitkänen [1, 2], Tuomo Nieminen[1], Antti Lipponen[2], Sini Isokääntä[1], Antti Arola[2], and Kari E. J. Lehtinen[1,2]

[1] Department of Applied Physics, University of Eastern Finland, Kuopio, Finland
[2] Finnish Meteorological Institute, Atmospheric Research Centre of Eastern Finland, Kuopio, Finland

Correspondence to: Santtu Mikkonen (santtu.mikkonen@uef.fi)

Abstract. Fitting a line on a scatterplot of two measured variables is considered as one of the simplest statistical procedures researchers can do. However, this simplicity is deceptive as the line fitting procedure is actually quite a complex problem. Atmospheric measurement data never comes without some measurement error. Too often, these errors are neglected when researchers are making inferences from their data.

To demonstrate the problem, we simulated datasets with different amounts of data and error, mimicking the dependence of atmospheric new particle formation rate ($J_{1.7}$) on sulphuric acid concentration (H_2SO_4). Both variables have substantial measurement error and thus they are good test variables for our study. We show that ordinary least squares (OLS) regression results in strongly biased slope values compared with six error-in-variables (EIV) regression methods (Deming, Principal component analysis, orthogonal, Bayesian EIV, and two different bivariate regression methods) known to take into account errors in the variables.

1 Introduction

Atmospheric measurement data never comes without some measurement error. Too often, these errors are neglected when researchers are making inferences based on their data. Describing the relationship between two variables typically involves making inferences in some more general context than was directly studied and if the relationship is ill formulated, the inference is not valid either. In some cases, the bias in analysis method is even given a physical meaning

When analysing dependencies of two or more measured variables, regression models are usually applied. A regression model can be linear or nonlinear, depending on the data. Standard regression models assume that the independent variables of the models have been measured without error and the models account only for errors in the dependent variables or responses. In cases where the measurements of the predictors contain error, estimating with standard methods, usually Ordinary Least Squares (OLS), do not tend to the true parameter values, not even asymptotically. In linear models, the coefficients are underestimated but in nonlinear models, the bias is likely to be more complicated (e.g. Schennach 2004). Measurement error needs to be taken into account especially when dealing with parameters with large errors. Thus, we chose such parameters as

our test variables in this study. Sulphuric acid (H_2SO_4) is known to affect strongly the formation rates (J) of aerosol particles (Kirkby et al., 2016; Kuang et al., 2008; Kulmala et al., 2006; Kürten et al., 2016; Metzger et al., 2010; Riccobono et al., 2014; Riipinen et al., 2007; Sihto et al., 2006; Spracklen et al., 2006). The relationship between J and H_2SO_4 is typically written as $\log_{10}(J) = \beta*\log_{10}(H_2SO_4)+\alpha$ (Seinfeld and Pandis, 2016). In addition, parameterizations based on the results from these fits have been implemented in global models, e.g. in (Dunne et al., 2016; Metzger et al., 2010; Spracklen et al., 2006), to estimate the effect of new particle formation on the global aerosol. Theoretically, the slope of this relationship is related to the amount of sulphuric acid molecules in the nucleating critical cluster in homogeneous nucleation, based on the first nucleation theorem (Vehkamäki, 2006)..

Some published results already show discrepancies regarding the J vs H_2SO_4 dependence. Kuang et al. (2008) used the unconstrained least squares method and obtained β=1.99 for the slope whereas Sihto et al. (2006) ended up with β=1.16 by using OLS from the same measurement campaign when analysing data from Hyytiälä in 2003. They had some differences in pre-treatment of data and used different time window but a notable proportion of this inconsistency is very likely due to different methods for making the fit. The problem in the relationship of H_2SO_4 and J has been acknowledged already in Paasonen et al. (2010) who noted that bivariate fitting method like presented in York et al. (2004) should be applied but could not be used due to the lack of proper error estimates for each quantity. They were not aware of the methods, which do not need to know the errors beforehand, but are using the estimated variances. We will here introduce appropriate tools for that.

Multiple attempts have been made to introduce methods using errors in predictor variables for the scientists applying regression-type analysis for their data; starting from Deming (1943). However, the traditional lest squares fitting still holds the position as the de facto line fitting method due to its simplicity. In atmospheric sciences, Cantrell (2008) drew attention to the method introduced by York (1966) and York et al. (2004) and listed multiple other methodological papers introducing similar methodology. Pitkänen et al. (2016) raised the problem into knowledge in remote sensing community and this study partly follows them and introduces multiple methods to take account the errors in predictors. Cheng and Riu (2006) studied methods with heteroscedastic errors whereas Wu and Yu (2018) approached the problem with measurement errors via weighted regression and applied some methods also used in our study.

Measurement errors in each variable of the model must be taken into account in the regression analysis by applying some errors-in-variables (EIV) regression. In this study, we compared OLS regression results to six different regression methods (Deming regression, Principal component analysis regression, orthogonal regression, Bayesian EIV regression and two different bivariate regression methods) known to be able to take into account errors in variables and provide (at least asymptotically) unbiased estimates. In this study, we will focus only on linear EIV methods but it is important to acknowledge that there also exist nonlinear methods described e.g. ORDPACK introduced in Boggs, Byrd, and Schnabel (1987) and implemented in Python SciPy and R (Boggs et al., 1989; Spiess, 2015). ORDPACK is somewhat improved version of orthogonal regression, as it minimizes the Mahalanobis distance from the data points to the regression line, instead of minimizing the sum of squares of the perpendicular distances, so that arbitrary covariance structures are acceptable and is specifically set up so that a user can specify measurement error variances and covariance point by point.

Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2018-1125
Manuscript under review for journal Atmos. Chem. Phys.
Discussion started: 11 December 2018

2 Materials and Methods

2.1 Data illustrating the phenomenon

In line fitting, data usually contain two types of error: natural error and measurement error. Measurement error is more generally understood, it is where measured values do not fully represent the true values of variable being measured. This also contains sampling error, e.g. in the case of H_2SO_4 measurement the sampled air in the measurement instrument is not representative sample of outside air. Natural error is that the true connection between the two variables is varying by some natural or physical cause e.g. certain amount of H_2SO_4 does not cause same number of new particles formed. In the analysis of measurement data, some amount of these errors are known or can be estimated, but some of it will usually remain unknown, which should be kept in mind when interpreting data. Even though the measurement error is taken into account, the regression fit may be biased due to natural error (Carroll and Ruppert, 1996).

The data used in this study consist of new particle formation rate at 1.7 nanometre size ($J_{1.7}$) and sulphuric acid (H_2SO_4) concentration simulated to mimic observations of pure sulphuric acid nucleation from CLOUD chamber in CERN (Kürten et al. 2016; https://home.cern/about/experiments/cloud) with matching expected value, variance and covariance structure. The chamber data at CERN is the best characterized and controlled set of new particle formation (NPF) experiments in the history of aerosol science so far. The Proton Synchrotron provides an artificial source of "cosmic rays" that simulates natural conditions of ionization between ground level and the stratosphere. The core is a large (volume 26m3) electro-polished stainless steel chamber with precise temperature control at any tropospheric temperature, precise delivery of selected gases and vapours and ultrapure humidified synthetic air. Existing data includes the most suspected candidates for atmospheric NPF, including sulphuric acid – ammonia – water (Kirkby et al., 2011), sulphuric acid – amine (Almeida et al., 2013) and ion induced organic nucleation (Kirkby et al., 2016).The actual nucleation of new particles occurs at slightly smaller size. After formation, they grow by condensation to reach the detection limit (1.7 nm) of the instrument and $J_{1.7}$ thus refers to the formation rate of particles as the instrument detects them. These variables were chosen because they are both known to have considerable measurement errors and their relationship is frequently under inference (Kirkby et al., 2016; Kürten et al., 2016; Riccobono et al., 2014; Tröstl et al., 2016) which makes them good illustrative variables for this study.

2.2 Regression methods

We made fits for the linear dependency of logarithms of the two study variables, such that the equation for the fit was given by

$$\boldsymbol{y = \beta_0 + \beta_1 x + \varepsilon} \quad (1)$$

where y represents $\log_{10}(J_{1.7})$, x is $\log_{10}(H_2SO_4)$, β:s are the coefficients estimated from the data and ε is the error term. In order to demonstrate the importance of taking into account the measurement errors in the regression analysis, we tested seven

different line-fitting methods. Ordinary Least Squares (OLS), not taking account the uncertainty in *x*-variable, and Deming regression (DR, Deming, 1943), Principal component analysis (PCA, Hotelling, 1957) regression, orthogonal regression (ODR, Boggs, Byrd, and Schnabel 1987), Bayesian EIV regression (Kaipio and Somersalo, 2005) and two different bivariate least squares methods by York *et al.*, (2004), and Francq and Govaerts (BLS, 2014), known to be able to take account errors in variables and provide (at least asymptotically) unbiased estimates. The differences between the methods comes from the criterion they minimize when calculating the coefficients and how they take account the measurement errors. The minimizing criteria for all methods are given in the supplement S1. Two of the methods, PCA and ODR, account only for the measurement variance, whereas Bayes EIV and York bivariate regression require known estimates for measurement errors. Though for Bayes EIV the error can be approximated with a distribution. DR and BLS can be applied with both, errors given by the user and measurement variance based errors. In this study, we applied measurement variance based errors for them.

3 Data simulation

In measured data the variables that are observed are not X and Y, but $(X+e_x)$ and $(Y+e_y)$, where e_x and e_y are the uncertainty in the measurements and the true X and Y cannot be exactly known. Thus, we chose to use simulated data, where we know the true X and Y, to illustrate how the different line fitting methods perform in different situations.

We simulated a dataset mimicking new particle formation rate ($J_{1.7}$) and sulphuric acid concentration (H_2SO_4) reported from CLOUD-chamber measurements in CERN. Both variables are known to have substantial measurement error and thus they are good test variables for our study. Additionally, the relationship of logarithms of these variables is quite often described with linear OLS regression and thus the inference may be flawed.

We generated one thousand random "true" H_2SO_4 concentration values assuming log-normal distribution with median $2.0*10^6$ and standard deviation $2.4*10^6$. The corresponding true $J_{1.7}$ was calculated using model $\log_{10}(J_{1.7}) = \beta*\log_{10}(\underline{H_2SO_4})+\alpha$ with the true slope β=3.3 and α=*-23*, both are realistic values presented by Kürten *et al.* (2016, Table 2 for the no added ammonia cases). The resulting $J_{1.7}$ mean was 0.11 and standard deviation was 0.50, similar to $J_{1.7}$ statistics in Kürten et al. (2016).

Simulated observations of the true H_2SO_4 were obtained by adding random errors $e_x = e_{rel,x}x + \sigma_{abs,x}$ that have a random absolute component $e_{abs,x} \sim normal(0,\sigma_{abs,x})$ and a random component relative to the observation x itself $e_{rel,x}x$, where $e_{rel,x} \sim normal(0,\sigma_{rel,x})$. Similar definitions apply for the true $J_{1.7}$, e_y, $\sigma_{abs,y}$ and $\sigma_{abs,y}$. The standard deviations of the measurement error components were chosen $\sigma_{abs,x} = 4*10^5$, $\sigma_{rel,x} = 0.3$, $\sigma_{abs,y} = 3*10^{-3}$, $\sigma_{rel,y} = 0.5$, which are subjective estimates based on measurement data.

Simulating the observations tends to generate infrequent extreme outlier observations from the infinite tails of the normal distribution. We discarded these outliers with an error larger than three times the combined standard uncertainty of the observation in order to remove the effect of outliers from the regression analysis. This signifies the quality control procedure in data analysis and it also improved the stability of our results between different simulations.

Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2018-1125
Manuscript under review for journal Atmos. Chem. Phys.
Discussion started: 11 December 2018

4 Results

Differences between the regression methods are illustrated with four different ways. First, by showing line fits on scatterplot of simulated data. Secondly, illustrating how the slopes change when the uncertainty in the measured variables increase, thirdly by showing the sensitivity of the fits on number of observations and finally showing how the fits are affected by adding outliers in the data.

Regression fits with all methods in use are shown in Figure 1. As we know that the "true slope" β_{true}=3.30 we can easily see how the methods perform. The worst performing method was OLS, with β_{ols}=1.55, which is roughly half of the β_{true}. The best performing methods with equal accuracy were ODR (β_{ODR}=3.27), Bayes EIV (β_{BEIV}=3.24) and BLS (β_{BLS}=3.22) whereas York (β_{York}=3.15), Deming (β_{DR}=2.95) and PCA (β_{PCA}=2.92) slightly underestimated the slope.

The sensitivity of the methods was first tested by varying the uncertainty in H_2SO_4 observations. We simulated six datasets with 1000 observations and with variating absolute and relative errors, listed in Table 1, and made fits with each method on all datasets separately. The performance of the methods is shown in Figure 2, with the results corresponding to Figure 1 are marked with black colour. It shows that when the uncertainty is smaller, the bias in OLS fit is smaller but the bias increases significantly when more uncertainty is added to data. Decrease in performance can also be seen with ODR, which is overestimating the slope, and PCA, DR and Bayes EIV, which all underestimate the slope. Bivariate methods, BLS and York, seem to be quite robust for increasing uncertainty, as the slopes are not changing considerably.

The sensitivity of methods on decreasing number of observations was tested by picking 100 random samples from the 1000 simulation dataset with n of 3, 5, 10, 20, 30, 50, 70, 100, 300 and 500 and making fits for all samples with all methods. The average slopes and their standard errors are shown in Figure 3. It is clear that when the number of observations is 10 or less, the variation in estimated slopes is considerably high. When $n \geq 20$ the average slopes stabilize close to their characteristic level, except for Bayes EIV, which needs more than 100 observations for that. The most sensitive methods for small n are Bayes EIV, ODR and PCA and thus they should not be applied for data with small n.

The sensitivity for outliers in predictor variable H_2SO_4 was tested with two different scenarios. First, the outliers were let to be randomly either high or low numbers. In the second scenario, outliers were allowed to be only high numbers, which is often the case in H_2SO_4 and aerosol concentration measurements as the lowest numbers are cleaned out from the data when they are smaller than the detection limit of the measurement instrument. Five cases with n=1000 were simulated with increasing number of outliers (0, 5, 10, 20, 100) and 10 repetitions of H_2SO_4 values with different set of outliers. Outliers were defined such that x_{obs}-x_{true}>3*combined standard uncertainty. The most sensitive methods for outliers in both scenarios were OLS and Bayes EIV. High number of outliers caused underestimation to PCA and DR, especially in high outlier case, and slight overestimation to BLS in random outlier case. York Bivariate and ODR were not affected at all in either case and BLS had only small variation between the 10 replicates in the estimated slope. We did not test how big number of outliers would break all of the methods as it might not be meaningful to interpret anymore data with more than 10% of outliers.

Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2018-1125
Manuscript under review for journal Atmos. Chem. Phys.
Discussion started: 11 December 2018

5 Conclusions

Simple linear regression can be used to answer some common questions, such as is *Y* related to *X* but if we are interested on the strength of the relationship then error-in-variance methods should be applied. There is no single correct method to make the fit, because the methods measure slightly different things about the data. The choice of method has to base on the properties of data and the specific research question. There are usually two types of error in the data: natural and measurement error. Even if the natural error in the data is not known, taking into account the measurement error improves the fit significantly. In addition, no matter how small the measurement error would be, it should be taken account because taking it into account will never lead to more biased estimator.

As a case study, we simulated a dataset mimicking the dependence of atmospheric new particle formation rate on sulphuric acid concentration. We introduced three major sources of uncertainty when doing inference from scatterplot data: increasing measurement error, number of data points and number of outliers. In Fig 1, we showed that in case of errors from real measurements of $J_{1.7}$ and H_2SO_4 four of the methods gave slopes close to "true" known value: BLS, York bivariate, Bayes EIV and ODR. Estimates from BLS and York bivariate remained stable even when the uncertainty in simulated H_2SO_4 was increased drastically in Fig 2. The main message to learn in Fig 3 is that with small numbers of observations all fit methods are highly uncertain. BLS showed out to be the most accurate with smallest sample sizes of 10 and less, ODR stabilized with 20 observations and York bivariate and Bayes EIV needed 100 or more data points to become accurate. After that, they approach the true value asymptotically, while the OLS slope, in contrast, converges towards an incorrect value. With the increasing number of outliers in Fig 4 ODR and York bivariate showed out to be the most stable ones, even with 10% of observations classified as outliers in both test cases. BLS remained stable in the case with only high outliers. Bayes EIV was the most sensitive to outliers with OLS.

From this, we can give a recommendation that if the uncertainty in predictor is known, York bivariate, or other method able to use known variances, should be applied. If the errors are not known, and they are estimated from data, BLS and ODR showed out to be the most robust in cases of increasing uncertainty and with high number of outliers. If the number of observations is less than 10, and the uncertainties are high, we recommend considering twice if a regression fit is appropriate at all. However, with our simulation tests BLS showed out to be the most robust with small data. Bayes EIV has significant advantages if the number of observations is high enough and there are not too many outliers, as it is able to estimate the errors in data with distributions.

Author contribution

SM prepared the manuscript with contributions from all co-authors. SM, MP and SI performed the formal analysis. MP simulated the data. SM, AA and KL formulated the original idea. SM, MP and AL developed and implemented the methodology. SM, MP, TN and AL were responsible of investigation and validation of data and methods.

Acknowledgments

This work was supported by The Nessling foundation and The Academy of Finland Centre of Excellence (grant no. 307331).

Competing interests

The authors declare that they have no conflict of interest.

Data availability: Simulated datasets used in the example analysis will be given as supplement upon publication.

6 References

Almeida, J., Schobesberger, S., Kürten, A., Ortega, I. K., Kupiainen-Määttä, O., Praplan, A. P., Adamov, A., Amorim, A., Bianchi, F., Breitenlechner, M., David, A., Dommen, J., Donahue, N. M., Downard, A., Dunne, E., Duplissy, J., Ehrhart, S., Flagan, R. C., Franchin, A., Guida, R., Hakala, J., Hansel, A., Heinritzi, M., Henschel, H., Jokinen, T., Junninen, H., Kajos, M., Kangasluoma, J., Keskinen, H., Kupc, A., Kurtén, T., Kvashin, A. N., Laaksonen, A., Lehtipalo, K., Leiminger, M., Leppä, J., Loukonen, V., Makhmutov, V., Mathot, S., McGrath, M. J., Nieminen, T., Olenius, T., Onnela, A., Petäjä, T., Riccobono, F., Riipinen, I., Rissanen, M., Rondo, L., Ruuskanen, T., Santos, F. D., Sarnela, N., Schallhart, S., Schnitzhofer, R., Seinfeld, J. H., Simon, M., Sipilä, M., Stozhkov, Y., Stratmann, F., Tomé, A., Tröstl, J., Tsagkogeorgas, G., Vaattovaara, P., Viisanen, Y., Virtanen, A., Vrtala, A., Wagner, P. E., Weingartner, E., Wex, H., Williamson, C., Wimmer, D., Ye, P., Yli-Juuti, T., Carslaw, K. S., Kulmala, M., Curtius, J., Baltensperger, U., Worsnop, D. R., Vehkamäki, H. and Kirkby, J.: Molecular understanding of sulphuric acid–amine particle nucleation in the atmosphere, Nature, 502(7471), 359–363, doi:10.1038/nature12663, 2013.

Boggs, P. T., Byrd, R. H. and Schnabel, R. B.: A Stable and Efficient Algorithm for Nonlinear Orthogonal Distance Regression, SIAM J. Sci. Stat. Comput., 8(6), 1052–1078, doi:10.1137/0908085, 1987.

Boggs, P. T., Donaldson, J. R., Byrd, R. H. and Schnabel, R. B.: Algorithm 676 ODRPACK: software for weighted orthogonal distance regression, ACM Trans. Math. Softw., 15(4), 348–364, doi:10.1145/76909.76913, 1989.

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(17), 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), 1–6, doi:10.1080/00031305.1996.10473533, 1996.

Cheng, C.-L. and Riu, J.: On Estimating Linear Relationships When Both Variables Are Subject to Heteroscedastic Measurement Errors, Technometrics, 48(4), 511–519, doi:10.1198/004017006000000237, 2006.

Deming, W. E.: Statistical adjustment of data, Wiley, New York., 1943.

Dunne, E. M., Gordon, H., Kürten, A., Almeida, J., Duplissy, J., Williamson, C., Ortega, I. K., Pringle, K. J., Adamov, A., Baltensperger, U., Barmet, P., Benduhn, F., Bianchi, F., Breitenlechner, M., Clarke, A., Curtius, J., Dommen, J.,

Donahue, N. M., Ehrhart, S., Flagan, R. C., Franchin, A., Guida, R., Hakala, J., Hansel, A., Heinritzi, M., Jokinen, T., Kangasluoma, J., Kirkby, J., Kulmala, M., Kupc, A., Lawler, M. J., Lehtipalo, K., Makhmutov, V., Mann, G., Mathot, S., Merikanto, J., Miettinen, P., Nenes, A., Onnela, A., Rap, A., Reddington, C. L. S., Riccobono, F., Richards, N. A. D., Rissanen, M. P., Rondo, L., Sarnela, N., Schobesberger, S., Sengupta, K., Simon, M., Sipilä, M., Smith, J. N., Stozkhov, Y., Tomé, A., Tröstl, J., Wagner, P. E., Wimmer, D., Winkler, P. M., Worsnop, D. R. and Carslaw, K. S.: Global atmospheric particle formation from CERN CLOUD measurements., Science, 354(6316), 1119–1124, doi:10.1126/science.aaf2649, 2016.

Francq, B. G. and Govaerts, B. B.: Measurement methods comparison with errors-in-variables regressions. From horizontal to vertical OLS regression, review and new perspectives, Chemom. Intell. Lab. Syst., 134, 123–139, doi:10.1016/j.chemolab.2014.03.006, 2014.

Hotelling, H.: The Relations of the Newer Multivariate Statistical Methods to Factor Analysis, Br. J. Stat. Psychol., 10(2), 69–79, doi:10.1111/j.2044-8317.1957.tb00179.x, 1957.

Kaipio, J. and Somersalo, E.: Statistical and Computational Inverse Problems, Springer-Verlag, New York., 2005.

Kirkby, J., Curtius, J., Almeida, J., Dunne, E., Duplissy, J., Ehrhart, S., Franchin, A., Gagné, S., Ickes, L., Kürten, A., Kupc, A., Metzger, A., Riccobono, F., Rondo, L., Schobesberger, S., Tsagkogeorgas, G., Wimmer, D., Amorim, A., Bianchi, F., Breitenlechner, M., David, A., Dommen, J., Downard, A., Ehn, M., Flagan, R. C., Haider, S., Hansel, A., Hauser, D., Jud, W., Junninen, H., Kreissl, F., Kvashin, A., Laaksonen, A., Lehtipalo, K., Lima, J., Lovejoy, E. R., Makhmutov, V., Mathot, S., Mikkilä, J., Minginette, P., Mogo, S., Nieminen, T., Onnela, A., Pereira, P., Petäjä, T., Schnitzhofer, R., Seinfeld, J. H., Sipilä, M., Stozhkov, Y., Stratmann, F., Tomé, A., Vanhanen, J., Viisanen, Y., Vrtala, A., Wagner, P. E., Walther, H., Weingartner, E., Wex, H., Winkler, P. M., Carslaw, K. S., Worsnop, D. R., Baltensperger, U. and Kulmala, M.: Role of sulphuric acid, ammonia and galactic cosmic rays in atmospheric aerosol nucleation, Nature, 476(7361), 429–433, doi:10.1038/nature10343, 2011.

Kirkby, J., Duplissy, J., Sengupta, K., Frege, C., Gordon, H., Williamson, C., Heinritzi, M., Simon, M., Yan, C., Almeida, J., Tröstl, J., Nieminen, T., Ortega, I. K., Wagner, R., Adamov, A., Amorim, A., Bernhammer, A.-K., Bianchi, F., Breitenlechner, M., Brilke, S., Chen, X., Craven, J., Dias, A., Ehrhart, S., Flagan, R. C., Franchin, A., Fuchs, C., Guida, R., Hakala, J., Hoyle, C. R., Jokinen, T., Junninen, H., Kangasluoma, J., Kim, J., Krapf, M., Kürten, A., Laaksonen, A., Lehtipalo, K., Makhmutov, V., Mathot, S., Molteni, U., Onnela, A., Peräkylä, O., Piel, F., Petäjä, T., Praplan, A. P., Pringle, K., Rap, A., Richards, N. A. D., Riipinen, I., Rissanen, M. P., Rondo, L., Sarnela, N., Schobesberger, S., Scott, C. E., Seinfeld, J. H., Sipilä, M., Steiner, G., Stozhkov, Y., Stratmann, F., Tomé, A., Virtanen, A., Vogel, A. L., Wagner, A. C., Wagner, P. E., Weingartner, E., Wimmer, D., Winkler, P. M., Ye, P., Zhang, X., Hansel, A., Dommen, J., Donahue, N. M., Worsnop, D. R., Baltensperger, U., Kulmala, M., Carslaw, K. S. and Curtius, J.: Ion-induced nucleation of pure biogenic particles, Nature, 533(7604), 521–526, doi:10.1038/nature17953, 2016.

Kuang, C., McMurry, P. H., McCormick, A. V. and Eisele, F. L.: Dependence of nucleation rates on sulfuric acid vapor concentration in diverse atmospheric locations, J. Geophys. Res., 113(D10), D10209, doi:10.1029/2007JD009253,

2008.

Kulmala, M., Lehtinen, K. E. J. and Laaksonen, A.: Cluster activation theory as an explanation of the linear dependence between formation rate of 3nm particles and sulphuric acid concentration, Atmos. Chem. Phys., 6(3), 787–793, doi:10.5194/acp-6-787-2006, 2006.

Kürten, A., Bianchi, F., Almeida, J., Kupiainen-Määttä, O., Dunne, E. M., Duplissy, J., Williamson, C., Barmet, P., Breitenlechner, M., Dommen, J., Donahue, N. M., Flagan, R. C., Franchin, A., Gordon, H., Hakala, J., Hansel, A., Heinritzi, M., Ickes, L., Jokinen, T., Kangasluoma, J., Kim, J., Kirkby, J., Kupc, A., Lehtipalo, K., Leiminger, M., Makhmutov, V., Onnela, A., Ortega, I. K., Petäjä, T., Praplan, A. P., Riccobono, F., Rissanen, M. P., Rondo, L., Schnitzhofer, R., Schobesberger, S., Smith, J. N., Steiner, G., Stozhkov, Y., Tomé, A., Tröstl, J., Tsagkogeorgas, G., Wagner, P. E., Wimmer, D., Ye, P., Baltensperger, U., Carslaw, K., Kulmala, M. and Curtius, J.: Experimental particle formation rates spanning tropospheric sulfuric acid and ammonia abundances, ion production rates, and temperatures, J. Geophys. Res. Atmos., 121(20), 12,377-12,400, doi:10.1002/2015JD023908, 2016.

Metzger, A., Verheggen, B., Dommen, J., Duplissy, J., Prevot, A. S. H., Weingartner, E., Riipinen, I., Kulmala, M., Spracklen, D. V, Carslaw, K. S. and Baltensperger, U.: Evidence for the role of organics in aerosol particle formation under atmospheric conditions., Proc. Natl. Acad. Sci. U. S. A., 107(15), 6646–51, doi:10.1073/pnas.0911330107, 2010.

Paasonen, P., Nieminen, T., Asmi, E., Manninen, H. E., Petäjä, T., Plass-Dülmer, C., Flentje, H., Birmili, W., Wiedensohler, A., Hõrrak, U., Metzger, A., Hamed, A., Laaksonen, A., Facchini, M. C., Kerminen, V. M. and Kulmala, M.: On the roles of sulphuric acid and low-volatility organic vapours in the initial steps of atmospheric new particle formation, Atmos. Chem. Phys., 10(22), 11223–11242, doi:10.5194/acp-10-11223-2010, 2010.

Pitkänen, M. R. A., Mikkonen, S., Lehtinen, K. E. J., Lipponen, A. and Arola, A.: Artificial bias typically neglected in comparisons of uncertain atmospheric data, Geophys. Res. Lett., 43(18), 10,003-10,011, doi:10.1002/2016GL070852, 2016.

Riccobono, F., Schobesberger, S., Scott, C. E., Dommen, J., Ortega, I. K., Rondo, L., Almeida, J., Amorim, A., Bianchi, F., Breitenlechner, M., David, A., Downard, A., Dunne, E. M., Duplissy, J., Ehrhart, S., Flagan, R. C., Franchin, A., Hansel, A., Junninen, H., Kajos, M., Keskinen, H., Kupc, A., Kürten, A., Kvashin, A. N., Laaksonen, A., Lehtipalo, K., Makhmutov, V., Mathot, S., Nieminen, T., Onnela, A., Petäjä, T., Praplan, A. P., Santos, F. D., Schallhart, S., Seinfeld, J. H., Sipilä, M., Spracklen, D. V, Stozhkov, Y., Stratmann, F., Tomé, A., Tsagkogeorgas, G., Vaattovaara, P., Viisanen, Y., Vrtala, A., Wagner, P. E., Weingartner, E., Wex, H., Wimmer, D., Carslaw, K. S., Curtius, J., Donahue, N. M., Kirkby, J., Kulmala, M., Worsnop, D. R. and Baltensperger, U.: Oxidation products of biogenic emissions contribute to nucleation of atmospheric particles., Science, 344(6185), 717–21, doi:10.1126/science.1243527, 2014.

Riipinen, I., Sihto, S.-L., Kulmala, M., Arnold, F., Dal Maso, M., Birmili, W., Saarnio, K., Teinilä, K., Kerminen, V.-M., Laaksonen, A. and Lehtinen, K. E. J.: Connections between atmospheric sulphuric acid and new particle formation during QUEST III–IV campaigns in Heidelberg and Hyytiälä, Atmos. Chem. Phys., 7(8), 1899–1914, doi:10.5194/acp-7-1899-2007, 2007.

Schennach, S. M.: Estimation of Nonlinear Models with Measurement Error, Econometrica, 72(1), 33–75, doi:10.1111/j.1468-0262.2004.00477.x, 2004.

Seinfeld, J. H. and Pandis, S. N.: Atmospheric chemistry and physics: From air pollution to climate change. [online] Available from: https://www.wiley.com/en-fi/Atmospheric+Chemistry+and+Physics:+From+Air+Pollution+to+Climate+Change,+3rd+Edition-p-9781118947401 (Accessed 26 September 2018), 2016.

Sihto, S.-L., Kulmala, M., Kerminen, V.-M., Dal Maso, M., Petäjä, T., Riipinen, I., Korhonen, H., Arnold, F., Janson, R., Boy, M., Laaksonen, A. and Lehtinen, K. E. J.: Atmospheric sulphuric acid and aerosol formation: implications from atmospheric measurements for nucleation and early growth mechanisms, Atmos. Chem. Phys., 6(12), 4079–4091, doi:10.5194/acp-6-4079-2006, 2006.

Spiess, A.: Orthogonal Nonlinear Least-Squares Regression in R, [online] Available from: https://cran.hafro.is/web/packages/onls/vignettes/onls.pdf (Accessed 17 July 2018), 2015.

Spracklen, D. V., Carslaw, K. S., Kulmala, M., Kerminen, V.-M., Mann, G. W. and Sihto, S.-L.: The contribution of boundary layer nucleation events to total particle concentrations on regional and global scales, Atmos. Chem. Phys., 6(12), 5631–5648, doi:10.5194/acp-6-5631-2006, 2006.

Tröstl, J., Chuang, W. K., Gordon, H., Heinritzi, M., Yan, C., Molteni, U., Ahlm, L., Frege, C., Bianchi, F., Wagner, R., Simon, M., Lehtipalo, K., Williamson, C., Craven, J. S., Duplissy, J., Adamov, A., Almeida, J., Bernhammer, A.-K., Breitenlechner, M., Brilke, S., Dias, A., Ehrhart, S., Flagan, R. C., Franchin, A., Fuchs, C., Guida, R., Gysel, M., Hansel, A., Hoyle, C. R., Jokinen, T., Junninen, H., Kangasluoma, J., Keskinen, H., Kim, J., Krapf, M., Kürten, A., Laaksonen, A., Lawler, M., Leiminger, M., Mathot, S., Möhler, O., Nieminen, T., Onnela, A., Petäjä, T., Piel, F. M., Miettinen, P., Rissanen, M. P., Rondo, L., Sarnela, N., Schobesberger, S., Sengupta, K., Sipilä, M., Smith, J. N., Steiner, G., Tomè, A., Virtanen, A., Wagner, A. C., Weingartner, E., Wimmer, D., Winkler, P. M., Ye, P., Carslaw, K. S., Curtius, J., Dommen, J., Kirkby, J., Kulmala, M., Riipinen, I., Worsnop, D. R., Donahue, N. M. and Baltensperger, U.: The role of low-volatility organic compounds in initial particle growth in the atmosphere, Nature, 533(7604), 527–531, doi:10.1038/nature18271, 2016.

Vehkamäki, H.: Classical nucleation theory in multicomponent systems, Springer-Verlag, Berlin/Heidelberg., 2006.

Wu, C. and Yu, J. Z.: Evaluation of linear regression techniques for atmospheric applications: the importance of appropriate weighting, Atmos. Meas. Tech., 11(2), 1233–1250, doi:10.5194/amt-11-1233-2018, 2018.

York, D.: Least-sqares fitting of a straight line, Can. J. Phys., 44(5), 1079–1086, doi:10.1139/p66-090, 1966.

York, D., Evensen, N. M., Martínez, M. L. and De Basabe Delgado, J.: Unified equations for the slope, intercept, and standard errors of the best straight line, Am. J. Phys., 72(3), 367–375, doi:10.1119/1.1632486, 2004.

Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2018-1125
Manuscript under review for journal Atmos. Chem. Phys.
Discussion started: 11 December 2018

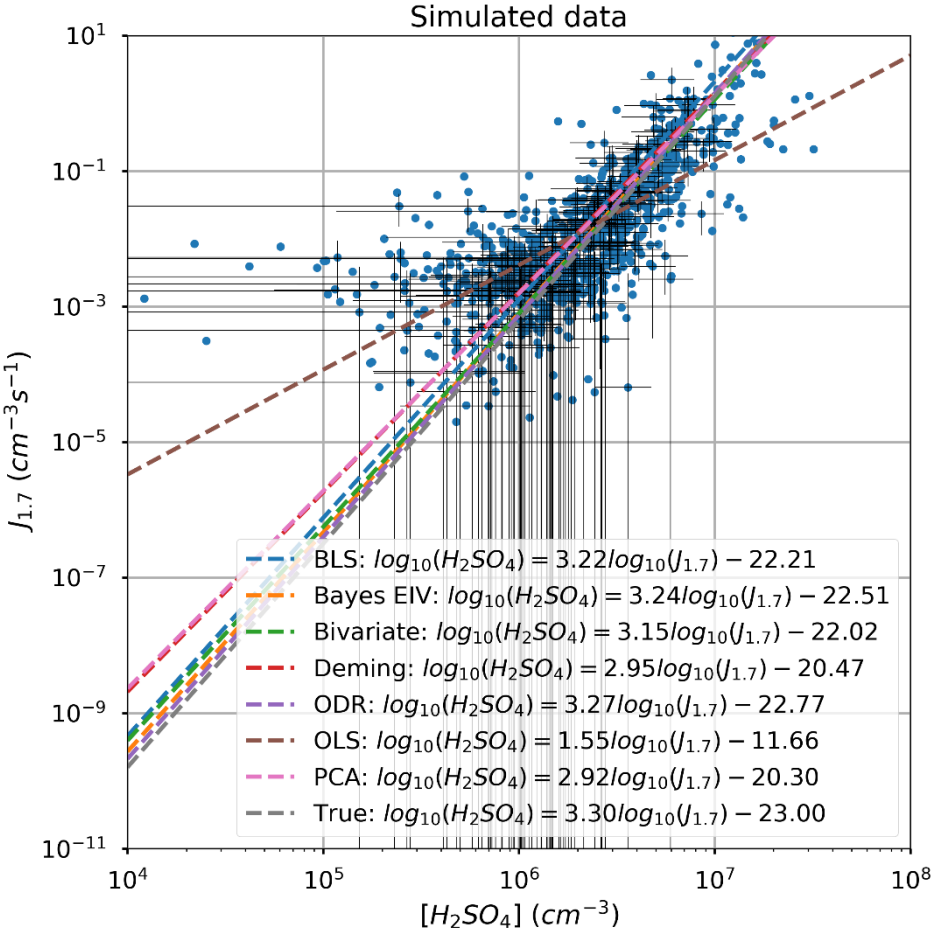

Figure 1. Regression lines fitted to the simulated data with all methods in comparison. Whiskers in data points refer to the measurement error used for simulation

Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2018-1125
Manuscript under review for journal Atmos. Chem. Phys.
Discussion started: 11 December 2018

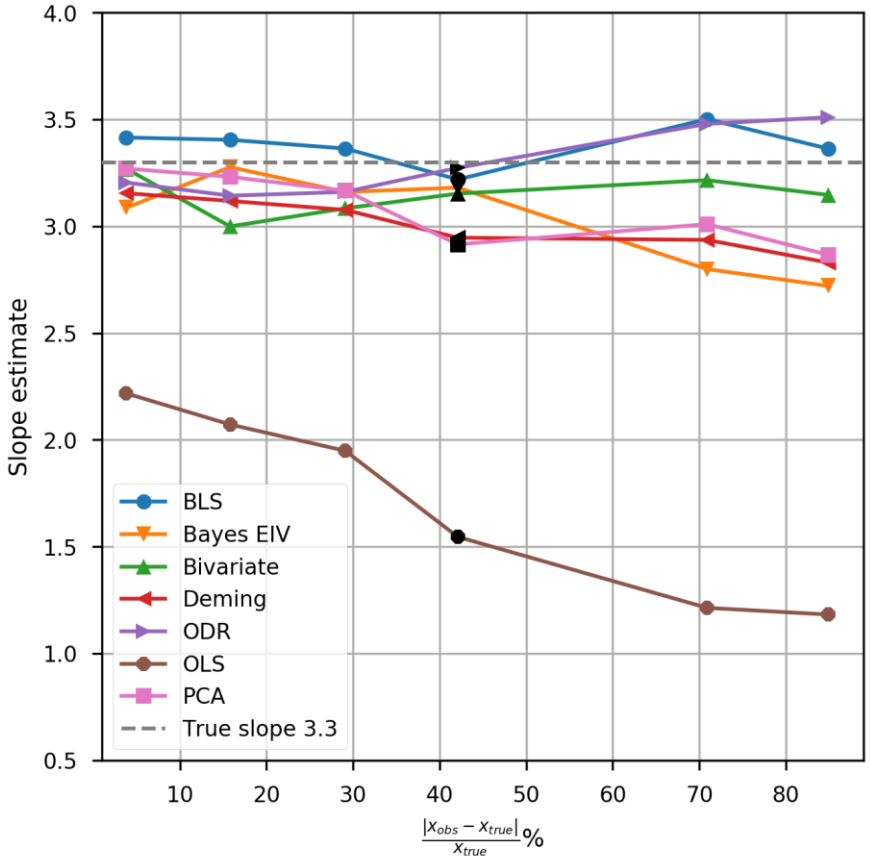

Figure 2. Sensitivity test for increasing uncertainty in simulated data. Black markers show the initial data set described in Section 3. Dashed line indicates the "true slope".

Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2018-1125
Manuscript under review for journal Atmos. Chem. Phys.
Discussion started: 11 December 2018

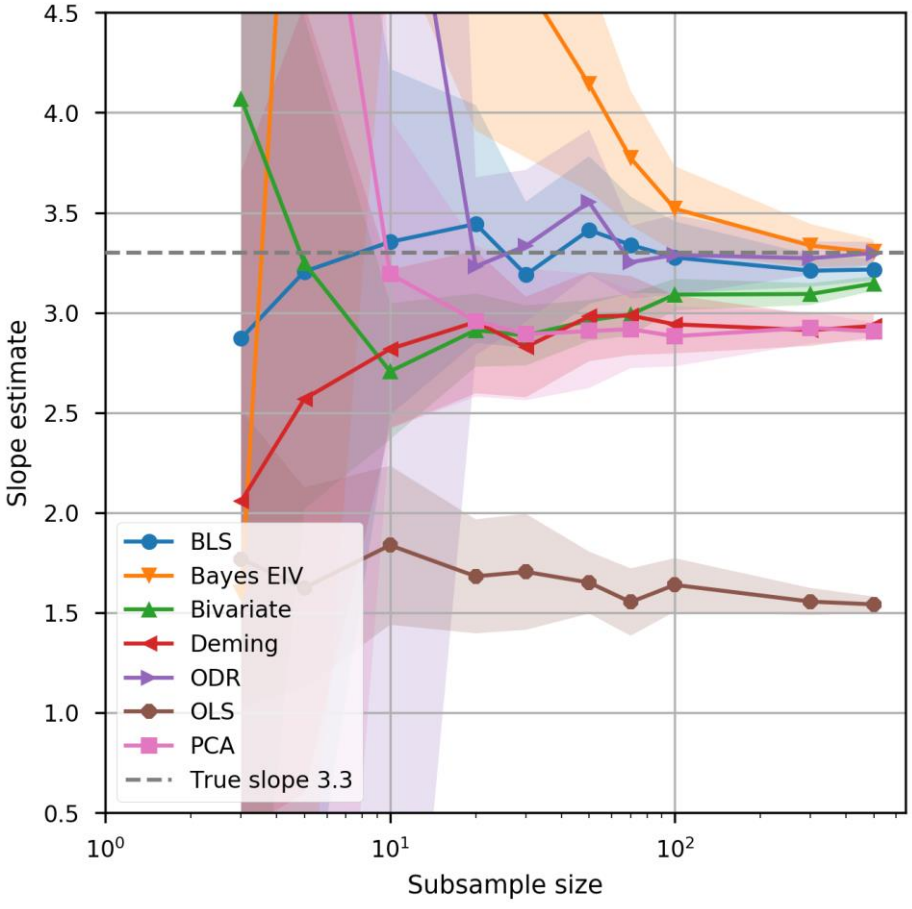

Figure 3. Effect of sample size on the uncertainty of different fits. Lines show the median and shading illustrates one standard deviation range of slope estimates for 40 repeated random samples. Dashed line indicates the "true slope".

Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2018-1125
Manuscript under review for journal Atmos. Chem. Phys.
Discussion started: 11 December 2018

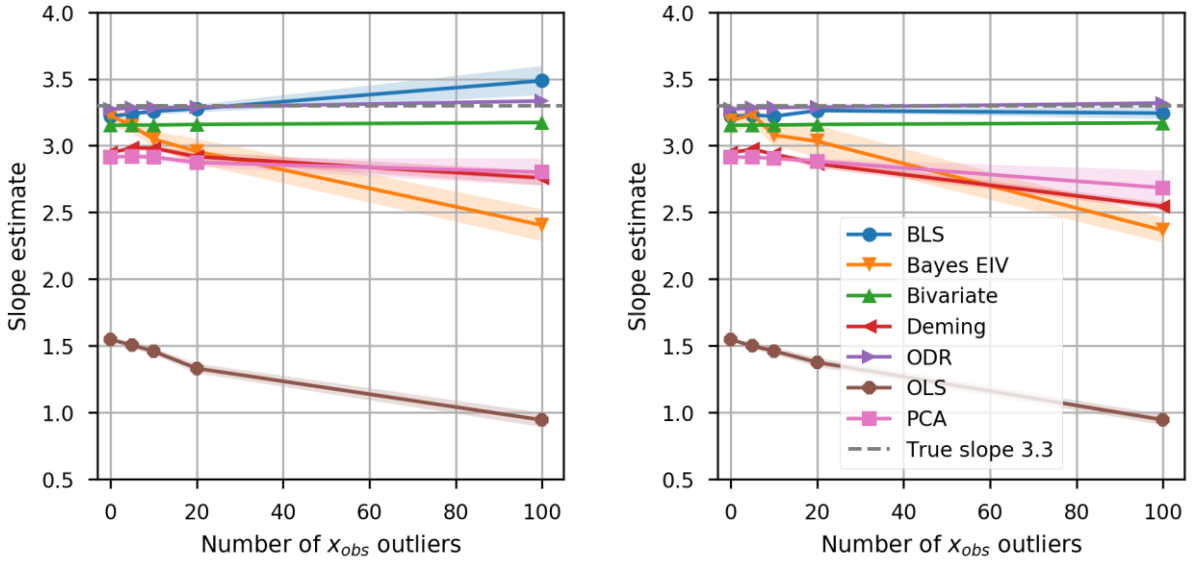

Figure 4. Effect of outliers in the data. Random outliers case on left panel and only high positives on right panel. Lines show the median and shading shows one standard deviation of slope estimates in ten repeated studies. Dashed line indicates the "true slope".

Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2018-1125
Manuscript under review for journal Atmos. Chem. Phys.
Discussion started: 11 December 2018

Atmospheric Chemistry and Physics Discussions

Table 1. The errors used in simulation for sensitivity test for increasing uncertainty

dataset	σ_{abs}	σ_{rel}	Ratio (= (σ_{rel} * $\overline{x_{obs}}$) / σ_{abs})
1	10^3	0.05	315.0
2	10^4	0.18	113.4
3	$7*10^4$	0.3	27.0
4	$4*10^5$	0.3	4.7
5	$6.5*10^5$	0.45	4.4
6	10^6	0.55	3.5