Anonymous Reviewer 1

Review of "Technical Note: Effects of Uncertainties and Number of Data points on Inference from Data – a Case Study on New Particle Formation" by Mikkonen et al.

This paper begins by discussion of using regression to infer aspects of observed data and goes on to describe the issues related to new particle formation rates. The heart of the paper involves generation of data purported to represent the logarithmic relationship between new particle formation rates and sulphuric acid concentrations. Several datasets are produced varying the amount of uncertainty, the sample size, and the number of outliers using seven regression procedures. From the regression fits, the paper makes recommendations as to when various procedures are appropriate.

This reviewer found the paper interesting and relevant for studies of atmospheric measurements, but in some cases the detail was not enough to assess the value of the results or the recommendations presented. The paper needs significant work before it is ready for publication. The authors should review the recommendations of the reviewers and make the needed changes. Perhaps the revised paper will be suitable for publication.

A: We thank the reviewer for the helpful comments and suggestions. Our answers to the concerns addressed are below.

General comments.

One of the key points of the paper was the inclusion of accurate estimates of errors in linear regression. This reviewer found the discussion of errors significantly lacking, and indeed including some incorrect statements. Within a measurement, there are two types of error: random and systematic.

Random errors can come from natural atmospheric fluctuations and instrument noise. Systematic errors can come from errors in calibration and loss of analyte in the inlet. This reviewer has never heard the term (nor could I find reference to) "natural error". One of the papers referenced (Carroll and Ruppert, 1996) also discusses "equation error", which refers to the errors associated with using an inappropriate form of a fitting equation. The paper needs a much more thorough description of errors, including introducing the symbols used later in the paper to describe errors.

A: We agree with the reviewer that the terminology with the errors needs to be clarified, as there are differences within substance areas. In statistical literature, systematic error is usually referred as "bias" and random error is divided in two parts, as reviewer indicated, which are caused by natural, stochastic, variation and the measurement itself (typically instrument or its user). The terminology is clarified in the revised manuscript, the section 2.1 now starts with text:

"Measurement data contains different types of errors. Usually, the errors are divided to two main class: random and systematic error. Systematic errors, commonly referred as bias, in experimental observations usually come from the measuring instruments. They may occur because there is something wrong with the instrument or its data handling system, or because the instrument is not used correctly by the operator. In line fitting, bias cannot be taken account but the random error may have different components, of which two are discussed here: natural error and measurement error. In addition, one should note the existence of equation error, discussed in Carroll and Ruppert (1996), which refers to using an inappropriate form of a fitting equation. 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 has stochastic variation by some natural or physical cause e.g. certain amount of H_2SO_4 does not cause same number of new particles formed."

The paper states on page 3, line 12 that the data used in this study are new particle formation rates and sulphuric acid concentrations. In fact, the data are simply calculations of two variables related by a linear relationship with noise added to represent random and systematic uncertainties (as done in other previous papers on linear regression). The data could represent any relation that is expected to be linear. The paper does not address nor answer any of the issues related to measurement or calculation of new particle formation rates except to say that one needs proper error estimates to perform regression on observed data, and that there are significant differences found depending on how data is handled. The reviewer finds this attempt to connect a linear regression paper to new particle formation without actually directly addressing the issue misleading. One solution would be to change the title, eliminating the part of about new particle formation, and to simply present new particle formation as one example of where error estimates are important for linear

regression. With the current title, the paper needs much more emphasis on the issues related to determining new particle formation using measurements and regression procedures.

A: The reviewer is correct that the sentence on page 3, line 12 in the original manuscript may give wrong impression on the data used, even though it stated that the data are "…concentrations simulated to mimic observations of…" and thus not claiming they are measured. We changed the beginning of the paragraph to form: "The data used in this study consist of simulated new particle formation rates at 1.7 nanometre size ($J_{1.7}$) and sulphuric acid (H_2SO_4) concentrations mimicking observations of pure sulphuric acid in nucleation experiments…"

New particle formation data was chosen as the basis of our simulated data because we think that with these kind of data inadequate analysis methods are often used regardless of the fact that the variables contain significant uncertainties. We agree with the reviewer that the data could be any set of numbers assumed to have linear relationship but to raise the awareness in the community we need to relate the simulations to well-known datatype. We added sentence on this to the end of chapter 2.2: "However, it should be kept in mind that the data could be any set of numbers assumed to have linear relationship but to raise the awareness in the research community we related the simulations to well-known datatype." However, in this type of short technical comment we do not want to take the attention from the analysis methods to specifics in NPF formation rate calculations, which are discussed in multiple papers and textbooks.

Several regression methods are used in the analysis, but the information about their use is superficial. For example, many of the methods are iterative. If proper convergence criteria are not set, then the results obtained are not useful. It is important to state the convergence criteria for each iterative method and state how it was determined that convergence was reached. For other methods, if there are adjustable parameters, these should also be discussed. Also, the software or program used for each of the methods should be given. If they are programs written in-house, it might be appropriate to make them available to the reader.

A: Information about which methods are iterative can be found in revised Supplement containing already the minimizing criteria. References on the software used is added to revised manuscript section 2.2: "The analysis for OLS and PCA were calculated with R-functions "lm" and "prcomp", respectively (R Core Team, 2018) DR was calculated with package deming

(Therneau, 2018) and BLS with package BivRegBLS (Francq and Berger, 2017) in R. The ODR based estimates were obtained using scipy.odr python package (Jones et al. 2001-), while the python package pystan (Stan Development Team, 2018) was used for calculating the Bayesian regression estimates. Finally, the York bivariate estimates were produced with a custom python implementation of the algorithm presented by York et al. (2004). " Convergence criteria in factory built functions are kept as default set by the writers of the software.

Specific Comments

It should be mentioned, perhaps in the introduction, that linear regression is appropriate when there are two measures of the same quantity (for example, by two different instruments) or when there are two measures that are related by a physical law (for example, the dependence of the logarithm of a rate coefficient on inverse temperature).

A: The reviewer is correct that this is important information but we want to believe that the readers of ACP are acquaint with basics of regression/line fitting.

Page 1, line 20. Suggest changing "comes" to "come" since strictly speaking "data" is plural (although often used singular).

A: Corrected as suggested

Page 1, line 22. Did not understand the "making inferences in some more general context than was directly studied". Suggest rewording or adding more information.

A: "inferences" changed to "deductions"

Page 1, line 23. Suggest "...the bias in the analysis method...". Sentence needs period.
Page 1, line 29. After "...coefficients are underestimated..." suggest adding a reference.
Page 1, line 29-30. Suggest "Measurement error needs to be taken into account, particularly when errors are large." Suggest removing "Thus, we chose such parameters as our test variables in this study." Suggest replacing it with "To demonstrate this point, we show the effects of large errors on linear regression in this study."
Page 2, line 1. Suggest "...known to strongly affect the formation...".

Page 2, line 3. Suggest "...between J and H_2SO_4 is typically assumed to be of the form: ...".

Page 2, line 6. Suggest "...formation on global aerosol amounts and characteristics. Theoretically in homogeneous nucleation, the slope of this relationship is related to the number of sulphuric acid molecules in the nucleating critical cluster, based on the...".

Page 2, line 9. Suggest "...results have shown discrepancies in the expected J vs. H_2SO_4 dependence."

Page 2, line 9-11. Suggest "Analysing data from Hyytiälä in 2003, Kuang et al. (2008) used an unconstrained least squares method and obtained β=1.99 for the slope, wheras Sihto et al. (2006) reported a value of 1.16 using OLS from the same field campaign."

Page 2, line 12. Suggest "...different time windows, but a significant proportion of this...".

Page 2, line 14. Suggest "...fitting method as presented in York..."

Page 2, line 15-16. Suggest "...of the methods that do not need to know the errors in advance, but instead made use of estimated variances."

Page 2, line 16. Suggest "Here, we present appropriate tools for using that approach."

Page 2, line 17. Suggest "...have been made to present methods accounting for errors in predictor variables for regression-type analysis, going back to Deming (1943)."

Page 2, line 19. Suggest "...due to its simplicity and common availability in frequently used software."

Page 2, line 20. Suggest "...methodological papers utilizing similar...".

Page 2, line 21. Suggest "...raised the awareness of the problem in the remote sensing...".

Page 2, line 22. Suggest "...follows their approach and introduces...".

Page 2, line 24. Suggest a different word that methods as it was used at the beginning of the sentence.

Page 2, line 25. Suggest "...in each variable must be taken into account using approaches called errorsin-variables (EIV) regression."

Page 2, line 30. Suggest remove "described.

A: All suggestions above are applied to the manuscript.

Page 2, line 31. Suggest "ORDPACK is a somewhat...".

Page 2, line 32. "Mahalanobis distance" is not a term most are familiar with. Might be worth a sentence and/or a reference to explain why it is different. Alternatively, perhaps leave out that detail.

A: Sentence corrected and comment on Mahalanobis distance removed

Page 3, Lines 4-25. In discussing new particle formation rates and the relationship to sulphuric acid concentrations, the authors might consider discussion the following subjects:

Are the errors in measurement of J and H_2SO_4 related?

What is known about other factors that might affect the relationship between J and H_2SO_4 (such as water vapor, temperature, pressure, etc.)?

A: In the simulated data the errors are not correlated but in the real measurements they might be. Even though the measurements are made with separate instruments, independent on each other, there might be come confounding factor effecting both of them at the same time. The factors listed by the reviewer are some of those. We added references to papers discussing these to the revised manuscript. Additionally, sentences referring to correlated error situation is added to this section: "In this study, we assume that the errors of the different variables are uncorrelated, but in some cases it has to be taken into account, as noted e.g. in Trefall and Nordö (1959) and Mandel (1984). The correlation between the errors of two variables, measured with separate instruments, independent on each other, like formation rate and H_2SO_4, may come e.g. from environmental variables affecting both of them at the same time. Factors affecting formation of sulphuric acid have been studied in various papers, e.g. in Weber et al. (1997) and Mikkonen et al. (2011). New particle formation rates, in turn, have been studied e.g. in Boy et al.(2008) and in Hamed et al. (2011) and similarities between affecting factors can be seen. In addition, factors like room temperature in measurement space and atmospheric pressure may affect to measurement instruments, thus causing additional error."

Page 3, Lines 4-11. See earlier comments about errors.

A: See comment above and corresponding modifications.

Page 3, line 12. Suggest "...particle formation rates at 1.7...".
Page 3, line 13. Suggest "...concentrations simulated...".
Page 3, line 13. Suggest"...pure sulphuric acid in nucleation experiments from the CLOUD..."

A: Corrections made as suggested for comments above

Page 3, line 14. Suggest "...with corresponding expected values, their variances, and the covariance structures."

A: Corrected as suggested

Page 3, line 15-16. It is clear you are proud of the accomplishments using CLOUD, but this reviewer suggests removing the sentence that begins "The chamber data at CERN...". Then, add CERN after "The" in the next sentence.

A: The data mimicking results from CLOUD was not used because we are proud of them but because we are concerned that many analyses on the data are made with methods not taking account the measurement errors. We added sentence on this to the end of section 2.1: "Additionally, many of the published papers on this topic do not describe how they are taking account the uncertainties in the analysis, which leaves a doubt that they are not treated properly."

Page 3, line 18. The word precise is used twice in this sentence, but it does not say how precise. Given the earlier comments this reviewer made about the lack of direct connection between this study and NPF studies, perhaps the details of CERN and NPF studies could be reduced or eliminated (lines 15-20). In this discussion, the connection between $J_{1.7}$ and H_2SO_4 concentration is not clearly demonstrated. Is it not true that the calculation involves corrections for condensation and (for some sizes) wall loss? Suggest being more complete or leaving out this part.
Page 3, line 19. If this sentence remains in the paper, need another word or more discussion of what is meant by "inference".

A: We added explanations to the use of the term "precise": "The core is a large (volume 26m3) electro-polished stainless steel chamber with temperature control (temperature stability better than 0.1 K) at any tropospheric temperature, precise delivery of selected gases (SO_2, O_3, NH_3, various organic compounds) and ultrapure humidified synthetic air, and very low gas-phase contaminant levels."

The connection between $J_{1.7}$ and H_2SO_4 is one of the key questions studied at CLOUD, and these studies utilize regression analyses. We chose to base our simulated datasets of $J_{1.7}$ and H_2SO_4 on data from CLOUD, because their well-controlled experiments make it possible to exclude other error sources than uncertainties on the $J_{1.7}$ and H_2SO_4. We added clarifications on the modified manuscript: "... and $J_{1.7}$ thus refers to the formation rate of particles as the instrument detects them, taking into account the known particle losses due to coagulation and deposition on the chamber walls. These variables were chosen because they are both known to have considerable measurement errors and their relationship is studied frequently using regression-based analyses"

Page 3, line 30. Change : to ´after β.

Page 4, line 12. Suggest "In measured data, the variables…"

Page 4, line 13. Suggest "...the measurements, and the true…." and "Thus, we use simulated data…"

Page 4, line 15. Suggest "...formation rates ($J_{1.7}$) and sulphuric acid concentrations…".

Page 4, line 20-21 and line 26. Suggest adding units to (molecules-cm^{-3}) to numbers.

Page 4, line 30. Suggest "This represents the quality…".

A: Corrections made as suggested for comments above

Page 4. Before starting the Results section, suggest some discussion of the fit methods, perhaps in the supplement. Suggest adding some basic introduction to the fit methods in the paper. This reviewer suggests testing the application of all the methods by testing with a known data set, such as Pearson´s data with York´s weights (York, 1966) whose fit parameters are known with very high accuracy.

A: We extended the descriptions on the methods to section 2.2 in the revised manuscript, as indicated in answers above. The reason for using simulated dataset in this study was that we would know exactly the expected value for slope and the errors. Thus using Pearson's data would not give that much additional value for the manuscript.

Page 5, line 8. It is not correct to say these methods had "equal accuracy" without stating the level of accuracy, in other words plus or minus an absolute level or plus or minus a percentage.

A: The sentence corrected to form: "The best performing methods with equal accuracy, i.e. within 2% range, were ODR (β_{ODR}=3.27), Bayes EIV (β_{BEIV}=3.24) and BLS (β_{BLS}=3.22), whereas York (β_{York}=3.15) was within 5% range, but Deming (β_{DR}=2.95) and PCA (β_{PCA}=2.92) slightly underestimated the slope."

Page 5, line 11. From the errors given in Table 1, show how the totals errors used in Figure 2 were calculated.

A: The relative errors (Fig. 2 horizontal axis values) were actually calculated from the simulated dataset values (as mentioned in Fig. 2 label) with $\frac{|x_{obs} - x_{true}|}{x_{true}}$ and not directly from the absolute and relative uncertainty values given in Table 1. That is, first each of the simulated data sets were generated as described in Section 3 and then the relative errors were calculated from the data itself.

Page 5, line 11. Suggest "...and with varying absolute and...".
Page 5, line 14. Suggest "...significantly as more uncertainty...".
Page 5, line 16. Suggest "...quite robust with increasing...".
Page 5, line 17. Suggest "...of methods to decreasing number...".
Page 5, line 20. Suggest "...estimated slopes can be very high."
Page 5, line 20. Suggest "...slopes stabilize close to their characteristic levels (within xx% for five methods) for large datasets."

A: Corrections made as suggested for comments above

Page 5, line 21. Suggest "...more than 100 observations."

A: Corrected as suggested

Page 5, line 22. It should be recognized that the number of points needed for a good fit depends on the uncertainties used. A few points will work fine if the uncertainties are small, while many more points are needed if uncertainties are large. This can perhaps be expressed at σ_x/x. Also, ensuring convergence is important for some of the methods (discussed above).

A: A sentence on the relationship of number of observations and the uncertainties was added at the end of this paragraph in the revised manuscript: "Though, it should be remembered that number of points needed for a good fit depends on the uncertainties in the data." Discussion on convergence was already added to method descriptions.

To get an accurate representation of the data, it is also helpful for the data to cover a wide range. The xdata in this study only cover the range from about 5 to 7 ($log_{10}[H_2SO_4]$). It would

be interesting for fits when the values covered a factor of 5 to 10, even if they are not realistic for actual atmospheric situations.

A: Wider range for X-data does not change the phenomenon, neither does varying the expected value of the "true slope". These were tested when preparing the manuscript and thus we are showing only atmospherically relevant numbers.

Page 5, line 24. This reviewer was not sure what is meant by "high and low numbers" and "high number" in this sentence. This needs more discussion and clarity for the reader to understand clearly what was done.

A: Unclear terms replaced by "high or low end of the distribution" and "large number" in the revised manuscript

Page 5, line 30. Suggest "...were not affected in either case...".

A: Corrected as suggested

Page 5, line 31. Suggest "We did not explore how large a number of outliers would be needed to seriously disrupt the fits for the various methods. We felt that it is likely not realistic to have situations with more than 10% outliers.

A: Corrected as suggested

Page 6, lines 2-4. This sentence needs rewording including improvement of the English to make it clear.

A: Sentence written in form: "Ordinary least squares regression can be used to answer some simple questions on data, such as is *Y* related to *X* but if we are interested on the strength of the relationship and the predictor variable *X* contains some error, then error-in-variables methods should be applied"

Page 6, line 4. Suggest "...of method should be based on the properties...".

A: Corrected as suggested

Page 6, lines 5-8. This should be reworked based on suggestions made above.

A: Definition of term “natural error” was inserted to the text as written above.

Page 6, line 11-12. It states that the fits are made with ″real″ data. This is not true. These are all synthetic data. It also says that four of the methods gave slopes close to the true value. Suggest a quantitative comparison: slopes are within 5% of the true value (or whatever is appropriate).

The methods are listed as good here are different than those listed in the Results section.

Suggest making this consistent.

A: The sentence means that the errors for the simulated data are taken from real measurements. The sentence is reformulated in the revised manuscript in order to avoid confusion, quantitative comparison is inserted as suggested and consistency with Results section is ensured. The new sentence is: “In Fig 1, we showed that in case of simulations where errors are taken from real measurements of $J_{1.7}$ and H_2SO_4 four of the methods gave slopes within 5% of the “true” known value: BLS, York bivariate, Bayes EIV and ODR.”

Page 6, line 14. It states that fits with small observations with all methods are highly uncertain. This does not agree with the earlier discussion and what is shown in Figure 3. Again, suggest quantitative comparisons and then statements about agreement (or lack of) that are also quantitative in this sentence and next few.

A: Uncertainty ranges of all methods in Fig. 3 are relatively high with small numbers of observations, even though average performance of BLS and, at some extent, York method are close to “true slope”

Page 6, line 15. Suggest ″BLS was the most accurate...″.

A: Corrected as suggested

Page 6, line 16. Statement does not agree with the that made in Results.

A: Statement in Results section was amended to be consistent with the conclusion

Page 6, line 18. Suggest "...number of outliers (Figure 4), ODR and the York bivariate methods were the most stable..."

A: Corrected as suggested

Page 6, line 20. Suggest "...sensitive to outliers after OLS."

A: Corrected as suggested

Page 6, line 22. The recommendations depend on the level of uncertainty. Suggest being more quantitative, in other words, something like "When errors (σ_x/x) are greater than 50%, then method x and y performed systematically better than methods w and z."

A: Recommendations quantified in the revised manuscript into form: "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 (relative error rE > 30% in Fig 2) and with high number of outliers. In our test data, BLS and ODR stayed stable up to rE >80% in Fig. 2 whereas DR and PCA started to be more uncertain when rE > 30% and Bayes EIV when rE>50%. "

Page 6, line 24. Suggest rewording "...we recommend considering twice...".

A: removed word "twice"

Page 6, line 25. Suggest "...robust with small numbers of data points." (Is this is what is meant?)
Page 6, line 32. Suggest "...were responsible for investigation...".

A: Corrections made as suggested for both comments

Anonymous reviewer #2

Interactive comment on "Technical note: Effects of Uncertainties and Number of Data points on Inference from Data – a Case Study on New Particle Formation" by Santtu Mikkonen et al.

"Technical note: Effects of Uncertainties and Number of Data points on Inference from Data – a Case Study on New Particle Formation" by Mikkonen et al. tests seven methods of linear regression on synthesized data. The resulting estimates of slope are compared among the methods and among various settings on uncertainty, sample volume and prescreening. This is a very nice study. It provides a concise reminder of how consequential the choice of linear regression method is. The example raised in the manuscript is easy to follow. While the subject is far from new, the atmospheric science has not seen a study with so many regression methods tested, as far as I know. The conclusions apply to A broad range of scientific analysis, in atmospheric science and beyond. I recommend publication. I only have minor suggestions to improve the readability.

Make the title more specific by including linear regression or line fitting, in addition to, or in place of, inference.

Shorten the second sentence in Introduction and the first sentence in Conclusions.

Treat "data" as either plural or singular, but not both, in the second paragraph of Section2.1.

Drop the "s" in "comes" in line 5, page 4.

Replace "on" with "in" in line 2, page 6.

Avoid placing the legend over the lines in Figure 3

A: We thank the reviewer on the positive feedback and helpful comments. Corrections to the revised manuscript are made as suggested and the title is changed in more specific form: "Technical note: Effects of Uncertainties and Number of Data points on Line Fitting - a Case Study on New Particle Formation"

Technical note: Effects of Uncertainties and Number of Data points on Line Fitting - 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
[§] Currently at: Institute for Atmospheric and Earth System Research, University of Helsinki, Helsinki, 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 come 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 deductions in some more general context than was directly studied. 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 (e.g. Carroll et al., 2006) but in nonlinear models, the bias is likely to be more complicated (e.g. Schennach

2004). Measurement error needs to be taken into account, particularly when errors are large.large Thus, we chose such parameters as our test variables in this study. Sulphuric acid (H_2SO_4) is known tostrongly affect 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 assumed to be in form $\log_{10}(J) = \beta*\log_{10}(H_2SO_4)+\alpha$ (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 global aerosol amounts and characteristics. Theoretically in homogeneous nucleation, the slope of this relationship is related to the number of sulphuric acid molecules in the nucleating critical cluster, based on the first nucleation theorem (Vehkamäki, 2006)..

Some published results have shown discrepancies in the expected J vs H_2SO_4 dependence. Analysing data from Hyytiälä in 2003, Kuang et al. (2008) used an unconstrained least squares method and obtained β=1.99 for the slope, wheras Sihto et al. (2006) reported a value of β=1.16 using OLS from the same field campaign. They had some differences in pre-treatment of data and used different time windows, but a significant 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 as 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 that do not need to know the errors in advance, but instead made use of estimated variances. Here, we present appropriate tools for using that approach.

Multiple attempts have been made to present methods accounting for errors in predictor variables for regression-type analysis, going back to Deming (1943). However, the traditional least squares fitting still holds the position as the de facto line fitting method due to its simplicity and common availability in frequently used software. 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 utilizing similar methodology. Pitkänen et al. (2016) raised the awareness of the problem in remote sensing community and this study partly follows their approach 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 techniques also used in our study.

Measurement errors in each variable must be taken into account using approaches called 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 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 a somewhat improved version of orthogonal regression 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.

2 Materials and Methods

2.1 Data illustrating the phenomenon

Measurement data contains different types of errors. Usually, the errors are divided to two main class: random and systematic error. Systematic errors, commonly referred as bias, in experimental observations usually come from the measuring instruments. They may occur because there is something wrong with the instrument or its data handling system, or because the instrument is not used correctly by the operator. In line fitting, bias cannot be taken account but the random error may have different components, of which two are discussed here: natural and measurement error. In addition, one should note the existence of equation error, discussed in Carroll and Ruppert (1996), which refers to using an inappropriate form of a fitting equation. 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 (e.g. due to losses of H_2SO_4 occurring in the sampling lines). Natural error is that the true connection between the two variables is has stochastic variation 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 unknown natural error. In this study, we assume that the errors of the different variables are uncorrelated, but in some cases it has to be taken into account, as noted e.g. in Trefall and Nordö (1959) and Mandel (1984). The correlation between the errors of two variables, measured with separate instruments, independent on each other, like formation rate and H_2SO_4, may come e.g. from environmental variables affecting both of them at the same time. Factors affecting formation of sulphuric acid have been studied in various papers, e.g. in Weber et al. (1997) and Mikkonen et al. (2011). New particle formation rates, in turn, have been studied e.g. in Boy et al.(2008) and in Hamed et al. (2011) and similarities between affecting factors can be seen. In addition, factors like room temperature in measurement space and atmospheric pressure may affect to measurement instruments, thus causing additional error.

The data used in this study consist of simulated new particle formation rates at 1.7 nanometre size ($J_{1.7}$) and sulphuric acid (H_2SO_4) concentrations mimicking observations of pure sulphuric acid in nucleation experiments from the CLOUD chamber in CERN (Kürten et al. 2016; https://home.cern/about/experiments/cloud) with corresponding expected values, their variances and covariance structures. The chamber data at CERNare 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 temperature control (temperature stability better than 0.1 K) at any tropospheric temperature, precise delivery of selected gases (SO_2, O_3, NH_3, various organic compounds) and ultrapure

humidified synthetic air, and very low gas-phase contaminant levels.~~,~~ Existing data include 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, taking into account the known particle losses due to coagulation and deposition on the chamber walls.~~,~~ These variables were chosen because they are both known to have considerable measurement errors and their relationship is studied frequently using regression-based analyses(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. Additionally, many of the published papers on this topic do not describe how they are taking account the uncertainties in the analysis, which leaves a doubt that they are not treated properly. However, it should be kept in mind that the data could be any set of numbers assumed to have linear relationship but to raise the awareness in the research community we related the simulations to well-known datatype.

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 orthogonal regression (ODR, Boggs, Byrd, and Schnabel 1987), Deming regression (DR, Deming, 1943), Principal component analysis (PCA, Hotelling, 1957) 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 come 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 S1but~~S1.~~but we give here the principles of the methods. OLS minimizes the squared distance of the observation and the fit line either in y or x direction, but not both at the same time, whereas ODR minimizes the sum of squared weighted orthogonal distances between each point and the line. DR was originallyis an orthogonal regression,alternative for ODR~~,~~ taking account the ratio of the error variances λ_{xy} of the variables, (in classical non-weighted ODR λ_{xy} =1) and it is the maximum likelihood estimate (MLE) for the model (1) when λ_{xy} is known. Idea of PCA is the same as in ODR but the estimation procedure is somewhat different as can be seen in S1. The bivariate algorithm by York et al 2004 provides a simple set of equations for iterating MLE of slope and intercept with weighted variables, which makes it similar to ODR in this case. However, ODR can consider linear scale uncertainties in logarithmic scale regression, unlike the York (2004) solution. In Bayes EIV, statistical models for the uncertainties in observed quantities are used and probability distributions for the line

slope and intercept are computed according to the Bayes' theorem. In this study, we computed Bayesian maximum a posteriori (MAP) estimates for the slope and intercept values. BLS takes into account errors and heteroscedasticity in both axes and thus is more advanced method than DR (under normality and homoscedasticity, BLS is exactly equivalent to DR). PCA accounts only for the measurement variance, whereas ODR, Bayes EIV and York bivariate regression require known estimates for measurement errors. Thought 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. The analysis for OLS and PCA were calculated with R-functions "lm" and "prcomp", respectively (R Core Team, 2018) DR was calculated with package deming (Therneau, 2018) and BLS with package BivRegBLS (Francq and Berger, 2017) in R. The ODR based estimates were obtained using scipy.odr python package (Jones et al., 2001), while the python package pystan (Stan Development Team, 2018) was used for calculating the Bayesian regression estimates. Finally, the York bivariate estimates were produced with a custom python implementation of the algorithm presented by York et al. (2004).

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 used 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 rates ($J_{1.7}$) and sulphuric acid concentrations (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 $2.0*10^6$ *(molecules cm^{-3})* and standard deviation $2.4*10^6$ *(molecules cm^{-3})*. The corresponding true $J_{1.7}$- was calculated using model $\log_{10}(J_{1.7}) = \beta*\log_{10}(H_2SO_4)+\alpha$ 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 represents the quality control procedure in data analysis and it also improved the stability of our results between different simulations.

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, i.e. within 2% range, were ODR (β_{ODR}=3.27), Bayes EIV (β_{BEIV}=3.24) and BLS (β_{BLS}=3.22), whereas York (β_{York}=3.15) was within 5% range, but 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 varying absolute and relative uncertainties, 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 as 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 with increasing uncertainty, as the slopes are not changing considerably.

The sensitivity of methods to 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 can be considerably high. When $n \geq 30$ the average slopes stabilized close to their characteristic level (within 5%), except for Bayes EIV and York bivariate, which needed more than 100 observations. 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. Though, it should be remembered that number of points needed for a good fit depends on the uncertainties in the data.

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 end of the distribution. In the second scenario, outliers were allowed to be only large numbers, which is often the case in H_2SO_4 and aerosol concentration measurements as the smallest 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 in either case and BLS had only small variation between the 10 replicates in the estimated slope. -We did not explore how large a number of outliers would be needed to seriously disrupt the fits for the various methods. We felt that it is likely not realistic to have situations with more than 10% outliers.

5 Conclusions

Ordinary least squares regression can be used to answer some simple questions on data, such as is Y related to X. However, if we are interested in the strength of the relationship and the predictor variable X contains some error, then error-in-variables 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 should be based on the properties of data and the specific research question. There are usually two types of error in the data: natural and measurement error, where natural error refers to stochastic variation in the environment. 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 simulations where errors are taken from real measurements of $J_{1.7}$ and H_2SO_4 four of the methods gave slopes within 5% of the "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 was 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 (Figure 4) ODR and York bivariate were 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 after 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 (relative error rE > 30% in Fig 2) and with high number of outliers. In our test data, BLS and ODR stayed stable up to rE >80% in Fig. 2 whereas DR and PCA started to be more uncertain when rE > 30% and Bayes EIV when rE>50%. If the number of observations is less than 10, and the uncertainties are high, we recommend considering if a regression fit is appropriate at all. However, with our simulation tests BLS showed out to be

the most robust with small numbers of data points. 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 for 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.

Boy, M., Karl, T., Turnipseed, A., Mauldin, R. L., Kosciuch, E., Greenberg, J., Rathbone, J., Smith, J., Held, A., Barsanti, K.,

Wehner, B., Bauer, S., Wiedensohler, A., Bonn, B., Kulmala, M. and Guenther, A.: New particle formation in the Front Range of the Colorado Rocky Mountains, Atmos. Chem. Phys., 8(6), 1577–1590, doi:10.5194/acp-8-1577-2008, 2008.

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.

Carroll, R. J., Ruppert, D., Stefanski, L. A. and Crainiceanu, C. M.: Measurement error in nonlinear models : a modern perspective., 2nd Editio., Chapman & Hall/CRC., 2006.

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 Berger, M.: BivRegBLS: Tolerance Intervals and Errors-in-Variables Regressions in Method Comparison Studies. R package version 1.0.0., [online] Available from: https://cran.r-project.org/package=BivRegBLS, 2017.

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.

Hamed, A., Korhonen, H., Sihto, S.-L., Joutsensaari, J., Järvinen, H., Petäjä, T., Arnold, F., Nieminen, T., Kulmala, M., Smith, J. N., Lehtinen, K. E. J. and Laaksonen, A.: The role of relative humidity in continental new particle formation, J. Geophys. Res., 116(D3), D03202, doi:10.1029/2010JD014186, 2011.

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.

Jones, E., Oliphant, T. and Peterson, P.: SciPy: Open Source Scientific Tools for Python, [online] Available from: http://www.scipy.org/, 2001.

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., 121(20), 12,377-12,400, doi:10.1002/2015JD023908, 2016.

Mandel, J.: Fitting Straight Lines When Both Variables are Subject to Error, J. Qual. Technol., 16(1), 1–14, doi:10.1080/00224065.1984.11978881, 1984.

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.

Mikkonen, S., Romakkaniemi, S., Smith, J. N., Korhonen, H., Petäjä, T., Plass-Duelmer, C., Boy, M., McMurry, P. H., Lehtinen, K. E. J., Joutsensaari, J., Hamed, A., Mauldin III, R. L., Birmili, W., Spindler, G., Arnold, F., Kulmala, M. and Laaksonen, A.: A statistical proxy for sulphuric acid concentration, Atmos. Chem. Phys., 11(21), 11319–11334, doi:10.5194/acp-11-11319-2011, 2011.

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.

R Core Team: R: A language and environment for statistical computing., [online] Available from: http://www.r-project.org, 2018.

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.

Stan Development Team: PyStan: the Python interface to Stan, Version 2.17.1.0., [online] Available from: http://mc-stan.org, 2018.

Therneau, T.: deming: Deming, Theil-Sen, Passing-Bablock and Total Least Squares Regression. R package version 1.4., [online] Available from: https://cran.r-project.org/package=deming, 2018.

Trefall, H. and Nordö, J.: On Systematic Errors in the Least Squares Regression Analysis, with Application to the Atmospheric Effects on the Cosmic Radiation, Tellus, 11(4), 467–477, doi:10.3402/tellusa.v11i4.9324, 1959.

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.

Weber, R. J., Marti, J. J., McMurry, P. H., Eisele, F. L., Tanner, D. J. and Jefferson, A.: Measurements of new particle formation and ultrafine particle growth rates at a clean continental site, J. Geophys. Res. Atmos., 102(D4), 4375–4385, doi:10.1029/96JD03656, 1997.

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.

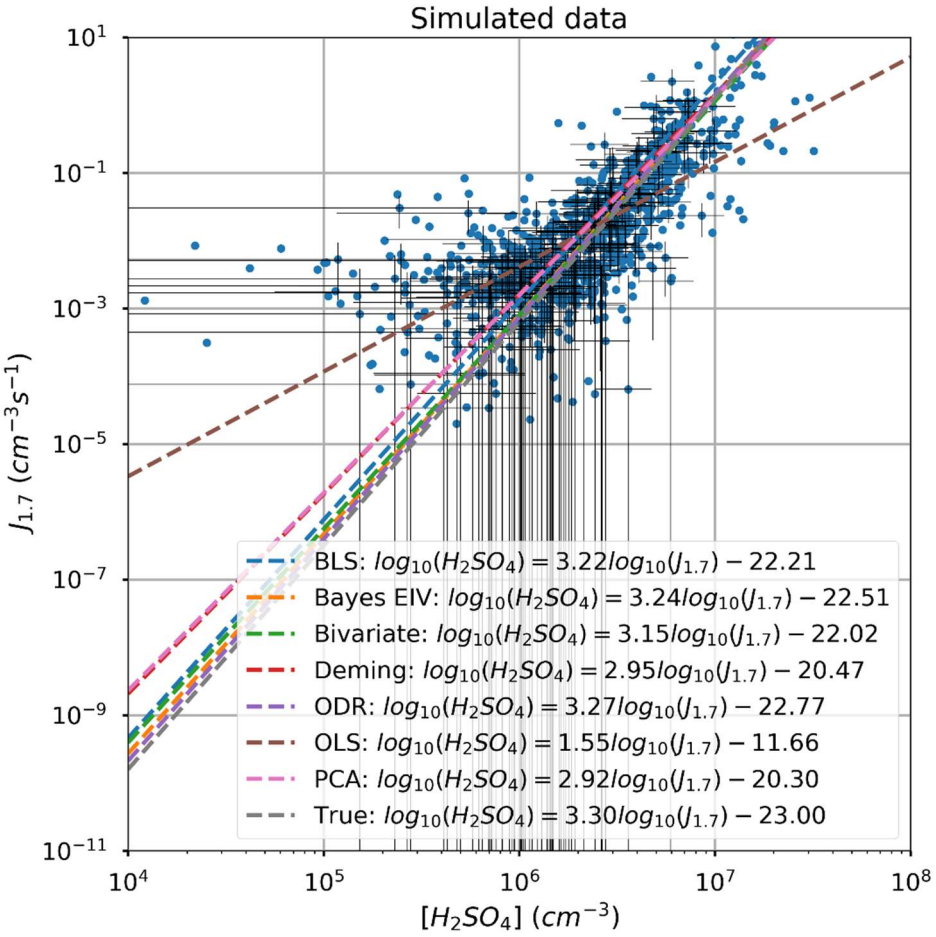

Figure 1. Regression lines fitted to the simulated data with all methods in comparison. Whiskers in data points refer to the measurement error used for simulation

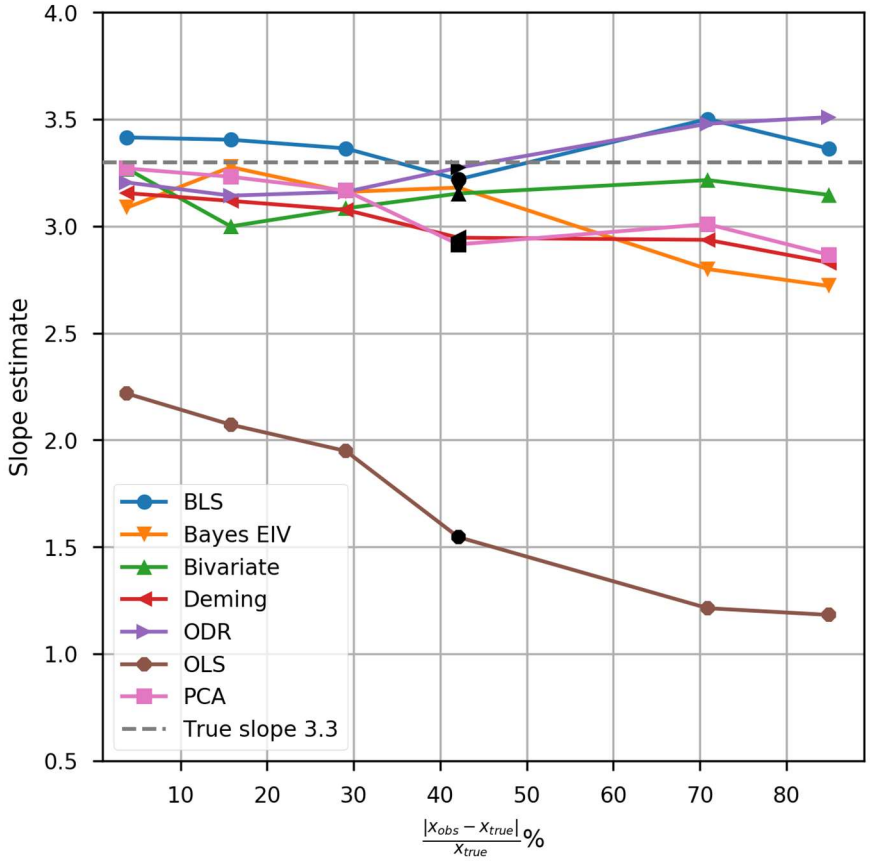

Figure 2. Sensitivity test for increasing uncertainty in simulated data. Black markers show the initial data set described in Section 3. Dashed line indicates the "true slope".

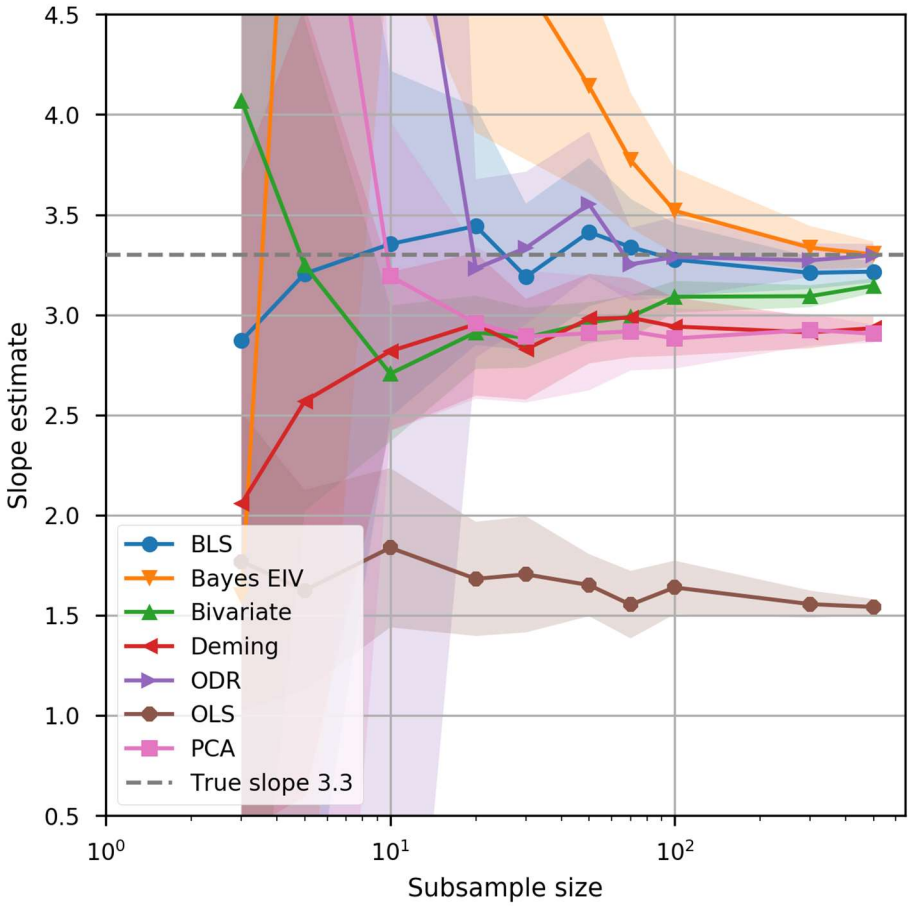

Figure 3. Effect of sample size on the uncertainty of different fits. Lines show the median and shading illustrates one standard deviation range of slope estimates for 40 repeated random samples. Dashed line indicates the "true slope".

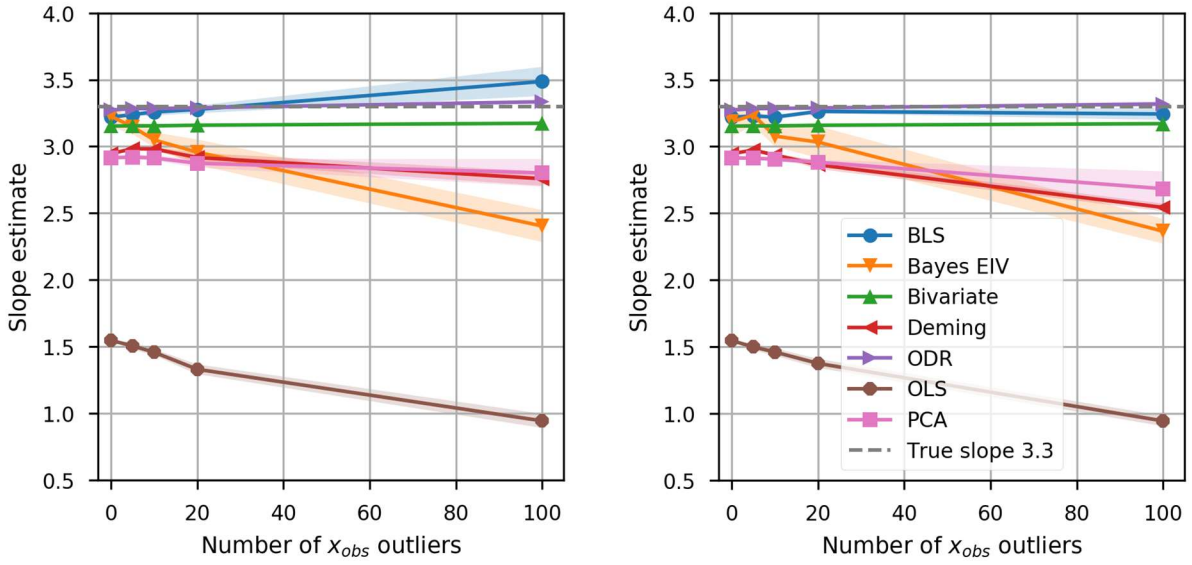

Figure 4. Effect of outliers in the data. Random outliers case on left panel and only high positives on right panel. Lines show the median and shading shows one standard deviation of slope estimates in ten repeated studies. Dashed line indicates the "true slope".

Table 1. The ~~errors~~ uncertainties 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