Response to Reviewer 1

We thank the reviewer for the helpful comments and suggestions. Below we address each of the reviewer's comments and indicate the requested changes to the manuscript. The reviewer's comments are marked in italic font and our response and changes to the manuscript are in plain text.

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.

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