Response to anonymous reviewer 2: Partridge et al., 2012.

The authors thank anonymous reviewer 2 for insightful comments on the manuscript. The reviewer provided several suggestions for improving the readability and quality of the manuscript. We have followed the suggestions in most cases, and our detailed response is outlined below.

Summary of main changes:

- We have modified the equations explaining the likelihood function used in the paper to make the methodology clearer. This is explained in detail in response to the individual review comments.
- We have extended the discussion in the paper with respect to the parameter sensitivity by adding an entirely new section in which we present results where the updraft velocity, mass accommodation coefficient, and surface tension are included as calibration parameters.
- From these new results we demonstrate that for the cloud parcel model employed and aerosol environments investigated, the updraft velocity is an extremely important parameter whereas the mass accommodation coefficient and surface tension are not.
- It is also demonstrated that the inclusion of these additional parameters does not affect the results or conclusions presented initially in the paper when we included only four calibration parameters.
- Figure 5 in the paper has been altered to present the results in a clearer manner.

Response to major comments

M1: *As in the review of the first part of this two part paper, I would have liked to see some discussion on design of experiments: how the simulated observations relate to possible real observations. Are there measuring devices available for such observations? Some parameters are seen to be uninformative with respect to the simulated observation. Would some reparameterization or different observational setup solve the problem?*

RM1: The parameters used in the inverse modeling were chosen with real world measurements in mind. For example, measured aerosol size distributions are routinely described in terms of

modal parameters (number, radius and GSD). Similarly, cloud droplet size distributions are measured in aircraft campaigns.

To explore the effect of number of calibration parameters we have added a section in the paper in which the updraft, mass accommodation coefficient and surface tension are also investigated. This section is titled: Inclusion of additional calibration parameters. The main results are not found to be significantly affected by the choice of number of perturbed parameters unless the aerosol Aitken lognormal parameters are also included.

We also performed additional simulations in which the interstitial aerosol was included in the calibration data and have briefly discussed this point. More detailed investigation of using additional observations is however out of the scope of this study, and is the subject of future research.

The same model parameters considered here are used in the third forthcoming paper in this series to describe measured aerosol physicochemical properties from the Marine Stratus/Stratocumulus Experiment (MASE II) campaign. Cloud droplet size distributions from the forward scattering spectrometer probe (FSSP) instrument were also measured during the same campaign and these are used in this forthcoming paper as the calibration data (along with the associated 'true' measurement error). This will allow us to assess model structural errors. These become apparent when the parameters are allowed to take on any value within their physically realistic ranges. The deviation from the inversely estimated parameters and their "measured" values is a useful diagnostic to assess the error in the model itself. At this point we are in a better position to assess whether any re-parameterisation of the adiabatic cloud parcel model is warranted by using real world observations.

Taking this into account, along with the fact that the cloud model setup used here is common to many similar studies in the literature, it seems prudent to retain the current modeling setup to keep a consistency with the final paper where the inverse modeling framework is applied to real world measurements.

M2: *Although the authors do not seem to agree, the use of same forward model to generate the data and solve the inverse problem will only provide information on the formal algorithmic flawlessness of the procedure. For relevant study of the underlying inverse problem, discretization, modelling error and ill-posedness of the problem, more comprehensive model for simulation of the "truth" is needed.*

RM2:

1.) This is the first time an MCMC algorithm has been coupled to a cloud model to explore cloud-aerosol interactions. For this reason we start with a model that is computationally efficient and introduce the framework to the audience before we couple the same model with real world measurements.

2.) If we used a more complex cloud model to generate our observations it is no longer possible to cross-compare the results directly as the aspect of the differences in model structure will severely complicate the analysis and throw people off the main objective of the paper.

3.) In addition, it is no longer possible to compare results as even though a synthetic truth is generated. The reason for this is that the parameters used in one model are not always directly related to model parameters from a second model. This then impedes us in benchmarking the approach. Also the output from a more complex model is much different (bulk values of droplet number compared to droplet size distribution if use 3D model for example). Using a different model for the 'truth' would complicate the analysis and the testing of the method and we believe this type of approach is outside the scope of the present work.

It is a good suggestion by the reviewer and we aim to pursue this approach in future research. However, investigating structural differences between models will undoubtedly be an involved process. Understanding the underlying algorithmic details and introducing the readers to the methodology in particular with regard to the cloud-aerosol inverse problem is a necessary first step so we can have confidence in applying this methodology to more complex problems.

Response to minor comments

1. *page 20053: line 3: In the term Markov chain Monte Carlo the word "chain" is not usually capitalized. This applies to the whole paper.*

This suggestion by the reviewer has been taken and the paper has been changed accordingly.

2. *page 20053: line 8: "... modelling framework is shown to successfully converge to ...". Framework converges? Maybe: "is shown to successfully estimate the correct calibration parameters".*

This suggestion by the reviewer has been taken and the paper has been changed accordingly.

3. *page 20059: line 10: "best values" is quite strong, and usually one refers to maximum a posteriori (MAP) instead of maximun likelihood when prior distribution is used explicitely.*

This suggestion by the reviewer has been taken so that (also called maximum likelihood) has been changed to maximum a posteriori density.

4. *page 20059: line 12: P is capitalized in P(theta|Y), but in the next line it is not.*

This was a typo; P is now not-capitalised throughout.

5. *page 20059: line 14: P(theta|Y) should be p(theta).*

We appreciate the reviewer's comments and have therefore clarified Eq. (1) by rewriting as

$$p(\theta|\boldsymbol{Y}) = \frac{p(\theta)p(\boldsymbol{Y}|\theta)}{p(\boldsymbol{Y})}$$

where $p(\theta)$ denotes the prior distribution of the parameters, and $L(\theta|\boldsymbol{Y}) \equiv p(\boldsymbol{Y}|\theta)$ signifies the likelihood function. We have also extended the description of the likelihood function.

6. *page 20059: line 15: Here you refer to the likelihood as objective function but later OF is defined as a function of model residuals.*

We appreciate the reviewer pointing out this, and have therefore altered Eq. 3 (Eq. 4 in updated paper) to show more clearly how the likelihood function of Eq. 1, $L(\theta|\boldsymbol{Y})$, is directly is related to the $OF(\theta)$ by adding Eq. 5.

7. *page 20059: line 17: Schoups and Vrugt (2009) is missing in the references.*

The reference has been added.

8. *page 20059: line 25: "before any data is collected" is not correct. Prior can, without any problems, depend on independent observations. Modelling is always an iterative process, as also discussed here, so both model and priors can be changed after iteration.*

We have modified the text at this location to clarify according to the reviewers comment.

9. *Page 20060: line 2: likelihood is (unnormalized) distribution of observations given the parameters. As the posterior is a product of prior and likelihood, the distributional form is important for the posterior. Least squares minimization corresponds to Gaussian likelihood, i.e. Gaussian observational error. From a statistical perspective, least squares is a consequence of Gaussian modelling of the errors.*

We have modified the text at this location to clarify according to the reviewers comment to: If we assume a standard Gaussian form of $L(\theta|\boldsymbol{Y})$, then the highest likelihood is typically found for those parameter values that provide the least squares fit to the experimental data.

10. *Page 20061: line 17: "powerful array of statistical measures" is not true in practice as convergence diagnostics can only diagnose non-convergence, not convergence (furthermore, their statistical power to detect that the chain has not converged can be low). I see no convergence diagnostics used in this paper.*

We have modified the text in the paper to clarify according to the reviewers comment and have also added text to clarify the use of the $\hat{R}$- statistic at the end of section 2.6.

11. *Page 20061: line 27: "detailed balance" and "ergodicity" are terms specific to Markov chain and MCMC literature and are possibly not very well understood by the readers of ACPD.*

We appreciate the reviewers comment, however, respectively choose to provide additional references to this point rather than explain in detail as we wish not to distract from the main focus of the paper. These references have been added to the paper.

12. *Page 20062: line 8: strictly speaking, the chain either is in the stationary distribution or not. In MCMC, stationary distribution is the limiting distribution so it is never reached, at least exactly.*

We have changed “to reach stationary” → “to travel to the posterior distribution”.

13. *Page 20066: line 2: "OF is simple least squares estimator". I think estimators and objective functions are different entities, although closely related. You could say that OF is the weighted sum-of-squares function. Estimator is what you get when you minimize the OF with respect to the parameter. Least squares and maximum likelihood estimators are the same for Gaussian likelihood and can even be said to correspond to the same OF, even if defined differently.*

We appreciate the reviewers comment and believe now that our modifications to the equations with respect to previous minor comments have now clarified this in the paper.

14. *Page 20066: line 10: The weights w_i which act as inverse variance of observational error are set to 1. But this value will certainly affect the size and shape of posteriors distributions as the more concentrated the posterior will be near the mode, the more Gaussian they tend to be, by simple linearization arguments. Later, observation error is set to be 10% Do you then change the weights accordingly?*

References to the weights have been removed from the text and the equations describing the fundamental statistical foundation of the method have been revised. The use of weights in the original version of the manuscript arose inadvertently; none of the experiments are run with a homoscedastic error. All the sensitivity experiments are run with a heteroscedastic error of 10% of the actual measured value. This value was used in Eq. 6. We employ a heteroscedastic error so the larger values are corrupted more -- the model has enough flexibility that it focuses more on the larger values than the small values. A heteroscedastic error is used as this was found to reflect the variability in real world observations. This is discussed further in our forthcoming part 3 of this paper series.

Slight differences in the results presented between the different versions of the paper are due to a slightly different formulation of the objective function in the first submission. This was based on Box and Tiao (1973); now we use the exact likelihood function with no proportionality form.

15. *Page 20067: line 11: The inclusion of Matlab command line is irrelevant here. You could just say that additive Gaussian noise is added to the observations with 10% standard deviation.*

We have modified the text at this location to clarify according to the reviewers comment.

16. *Page 20067: line 6: The word "corrupt" seems out of contents. Every measurement is bound to have some noise. If an observation is "corrupt", it typically is an outlier of some sort.*

We appreciate the reviewer's comments, but respectively state that "corrupt" has been used elsewhere in the literature. However, to make this clearer corrupt has been changed to perturb throughout the text.

17. *Page 20068: line 20: Performance of MCMC algorithm.*

This section explains how DREAM works for error free observations and for observations with 10% error added. The authors show that the algorithm is "performing" in the sense that is works correctly.

For a real performance study, one would like to see comparisons to alternative methods, such as plain Metropolis-Hastings using numerical Jacobian for construction proposal distribution, for example, some timings, estimates of Monte Carlo errors of the chain estimates or estimate of the efficient number of simulations. There is mild nonlinearity in the pairwise correlation figures (e.g., Fig.9) but no sign of multi modality. Also, the number of parameters is only four. Would it be possible to solve this problem with more standard MCMC algorithms, even more efficiently?

We respectfully beg to differ in opinion with this reviewer. In the first place, several previous publications have shown that the DREAM package outperforms existing RWM, DRAM, AM and AP algorithms, in particular when confronted with high-dimensionality and multi-modality. Indeed, the present case study is rather simple, but still DREAM will exhibit excellent performance (as demonstrated in Figure 5). In the second place, DREAM is an exact sampler and therefore converges to the exact posterior distribution; the posterior distributions reported herein should be exact and will not change if another MCMC sampler is used. Thirdly, the goal of this paper is not to compare different MCMC algorithms – this has already been done in our previous work for a range of different problems involving anywhere between 2 – 300 dimensions.

18. *In Figs 3 and 4 the MCMC chain plots show the upper/lower prior limits are reached for several of the parameters. Either the observations are not informative or prior specifications are too restricted. It is typical that experts, not accustomed to provide multidimensional priors, tend to think in terms of one dimensional conditional distributions instead of marginal distributions when defining limits, and thus give too limited bounds.*

We appreciate this comment of the reviewer, but the prior ranges of the parameter have been carefully and rather non-conservatively selected based on extensive literature search. This is highlighted in section 2.4 of the paper, as well as in P11. Larger ranges of the individual parameters cannot be justified, and are simply physically unrealistic.

19. *Page 20070: line 18: Only 20% of the model simulations are used in analysis. This seems rather inefficient. Do you recommend this as a general rule for DREAM, or have you used some diagnostics to infer this percentage?*

For simplicity we just used the last 20% of our simulations. In principle, we could take all those simulations for which the R-statistic is smaller than 1.2; but resort to the last 20% given that we have done enough simulations.

20. *Page 20070: line 25: It is not clear from the text what is meant by relative sensitivity. Caption in Fig. 7 suggests that the prior range is used as scale. This is not stated in the text. This sensitivity, as said, depends solely on the relative choice of the prior bounds. And also, as discussed above, it depends on the weights w_i used in the likelihood "OF".*

In section 3.3.1 we stated that “In order to confirm these preliminary indications and see the true relative sensitivity between different calibration parameters we normalise the posterior ranges by the prior ranges for each individual parameter”.

We have reworded this section in the paper to make it clearer.

21. *Page 20080: line 18: maybe you should re-think the points given as advantages; are these really the main advantages. In the third item, problems are given as advantages. About the fourth item: are you running the code in parallel, or is this just a possible advantage.*

We agree with the reviewers advice and have changed the word advantages to merits so that the bullet points are more meaningful, and have also moved the items not suited to this heading to the following heading now termed considerations.

22. *Page 20081: line 8: limitations of what? Why is the second item "a limitation"?*

We have changed the word limitations to considerations so that the bullet points are more meaningful.

23. *Page 20081: line 22: Why do you say that introducing more prior would produce more "confident" sensitivity estimates? Confident in what sense? Wouldn't the posteriors then include more information from the prior and less from the observations.*

The relative sensitivities are of course directly dependent on the choice of the prior ranges. We made a particular effort to specify these in as realistic a manner as possible through a comprehensive literature search. We therefore feel confident that the results are statistically meaningful; however, in the case of non-identifiable parameters it would not matter anyway; whether you use a range between 1-2 or 1 – 10; the posterior extends both these ranges and in both cases the sensitivity is 0%. We have modified the text accordingly to make out point clearer.

24. *Page 20082: line 23 "the objective being ..." this sentence is very hard to understand. What model input parameters? Do you mean a priori parameter values? What is a "measurement estimate"? What are "associated observations"?*

We agree with the reviewer that the wording of this sentence needed improving and have done so.