Response to discussion-stage referee comments for the paper "Identification of molecular cluster evaporation rates, enthalpies and entropies by Monte Carlo method"

immediate

June 27, 2020

1 Overview

In this document we respond to the referee comments for the paper "Identification of molecular cluster evaporation rates, enthalpies and entropies by Monte Carlo method". These comments were provided at the public discussion stage of the review process for publication in Atmospheric Chemistry and Physics.

In Section 2 we list each of Referee's comments. We also include our comment-by-comment responses. Each of the referee's comments are denoted with "**C**" and our responses to the referee's comments are denoted with "**R**".

We thank the referee for his/her time, thoughtfulness, and feedback. All the remarks and suggestions for our paper have been very helpful.

2 Referee 1 comments and our responses

Referee 1's summary: This manuscript applies Markov Chain Monte Carlo method to estimate cluster evaporation rates and cluster thermodynamic parameters such as formation enthalpies and entropies while taking

collision rates from kinetic gas theory. Cluster evaporation rates were estimated from two data sets: steady-state and transient data. While the transient data can improve the estimates of the evaporation rates compared to the steady state data, neither of them can be satisfied from both magnitude and the marginal posterior distributions of the rates. Cluster formation enthalpies and entropies were then estimated from steady-state cluster concentrations at two temperatures (278 and 292 K) and the cluster evaporation rates were inversed from the cluster Gibbs free energies (determined by enthalpies and entropies). It turns out that the evaporation rates were greatly improved in terms of variation and the probability distributions except for clusters containing both 5 sulfuric acid and 5 ammonia. Since cluster evaporation rate is an essential parameter that controls cluster growth, this parameter ought to be accurately determined in order to understand atmospheric nucleation. The scientific questions are worthy exploring and are important topics in atmospheric research. However, several major issues need to be fully resolved before the manuscript is considered for publication in this journal.

1. **C:** Section 2: the way the authors describe simulation methods is hard to understand. It seems that the authors wrote paragraphs in casual ways, in particular, when describing MCMC simulations, it is very hard to follow the logic. It is suggested that the authors use more plain languages and better logic to rearrange section 2 in order for readers to understand the methods and data sets the authors used or generated.

   R: We have cleaned up the wording in several places in Section 2. Below are the changes we have made.

   - In section 2 just before subsection 2.1, we added "In this section we describe the methods used to create synthetic cluster concentration data sets. We also explain the Monte Carlo type algorithms used to estimate the cluster evaporation rates from the data sets."
   - In line 93, added "particle" before the word cluster.
   - In line 102 we replace "(see the Table 2)" with the sentence "See Table 2 for the summary of ammonia mixing ratio and the source of sulphuric acid monomer used for the ACDC simulations".
   - Starting from line 103, rewrote the paragraph to read: "First, we computed the collision rates using the Eq. A3 from kinetic gas theory. Then, we were using these values for the collision rates along with Eq. A4 and the Gibbs free energies computed from Eq. A5 to obtain the evaporation rates. Note that to compute

the Gibbs free energies, we substituted the values for cluster formation enthalpies and entropies given by Olenius et al. (2013b) into Eq. A5. Additionally, we consider the losses on the CLOUD chamber walls which depend on the cluster size computed with Eq A5 (see Kürten (2015)) and a dilution loss of $S = 9.6 \times 10^{-5}$ s^{-1}. These values for the rates and losses were substituted into the ACDC algorithm (see McGrath et al. (2012)), which simulates the time evolution of molecular cluster concentrations. The ACDC code computes the first-order non-linear, ordinary differential system of cluster concentrations as given by Eq. A1. We then integrate the system produced by ACDC using the Fortran ordinary differential equation solver VODE (N. Brown et al. (1989)). A detailed description of this strategy for solving the forward-problem of finding the cluster concentration rates from Eq. A1 was published in McGrath et al. (2012). To reproduce the experimental conditions as realistically as possible, each simulation was initialized with non-zero concentration of ammonia monomer and no sulphuric acid. The source of sulphuric acid monomer was supplied at a constant rate.

The above method we used for producing synthetic concentration rates is similar to the one described in Kupiainen-Määttä (2016). We note that unlike Kupiainen-Määttä (2016), in this paper, our particle system is considered at various temperatures."

- In line 110, we changed the first sentence to "Using the above algorithm, model configuration and parameters, we generated two data sets."
- In line 111, we changed the sentence "The maximum time we run is 60 minutes in the above model configurations" to "The maximum time we run is 60 minutes from beginning of the simulation, in the above model configurations"
- In line 112, we reformulated the sentence to clarify how the time-dependent synthetic data were generated: "In this case, it is assumed that the concentrations for all the clusters are measured under constant temperature with time resolution comprising 1.5 minutes, which comprises overall 41 time-dependent concentration data for each of the cluster types i measured from beginning to the end of each ACDC simulation, before the system has attained a steady state."
- In line 114, we added at the end of the sentence

- In line 127, we added the sentence "Now we describe how we estimate the evaporation rates from the noisy synthetic data sets obtained by the method described in Section 2.1. We first give a general overview of the basic Metropolis algorithm (Metropolis (1953)), then describe a modification of the algorithm we implemented in this study, and finally, in Section 2.2.3 we apply this general framework to each of our study cases."
- We added section 'The Metropolis algorithm' restructured the Section 2.2 into three sub-sections,
- We changed the sentences starting from line 129 to read The objective of MCMC in parameter estimation is to identify all the possible parameter values which yield the best fit with the experimental data. Unlike optimization algorithms that produce one best combination of parameter values, the in the MCMC procedure all the most-probable combinations of parameter values are estimated given the data. To obtain these combinations, the values of parameters are generated and stored into the MCMC "chain". The MCMC chain will converges to the distribution containing all the most-likely combinations of parameter values as a number of sampled parameter sets (i.e., the chain length) increases. The distribution formed from the chain approximates a posterior probability density function which gives the likelihood of observing each of the parameters given the concentration data.
- To make the MCMC workflow more logical, we rearranged the remaining content of Section 2.2 into 3 subsections: "The Metropolis algorithm" (Section 2.2.1), "The DRAM algorithm" (Section 2.2.2) and "The overview of the MCMC runs" (Section 2.2.3). The fist section explains the basic Metropolis algorithm, the second section gives a detailed description of the Delayed Rejection Adaptive Metropolis algorithm used in the present study, the last subsection explains the domain restrictions for sampled parameters and parameter representation of the evaporation rates.
- After the line 132 We added subsection with the caption 'The Metropolis algorithm'.
- Starting with line 133, we wrote the subsection describing the basic Metropolis algorithm in application to our simulation: "First, a prior distribution for the parameter values $\boldsymbol{\theta}$ (represented in array form) is chosen and set to be the proposed "true" distribution from which possible parameters are sampled. The prior is typi-

cally selected based on the previous knowledge for the parameter values. Then an initial guess for parameter values (denoted as θ_0 or θ_{old}) is selected from the prior distribution.

Starting from the initial guess, the algorithm samples candidate parameter values (denoted as θ_{new}) from a proposal distribution centred at the previous point (denoted as $q(\theta_{\text{old}}, \theta_{\text{new}})$). The proposal density $q(\theta_{\text{old}}, \theta_{\text{new}})$ is symmetric, which means that the probability of step taken from the 'old' θ_{old} to the 'new' point θ_{new} is same as the probability of the reverse step ($q(\theta_{\text{old}}, \theta_{\text{new}}) = q(\theta_{\text{new}}, \theta_{\text{old}})$).

Then the candidate point $\boldsymbol{\theta}_{\text{new}}$ is either accepted or rejected, according to the least-squares fit of the output to the data, which measures the difference between the modelled $\mathbf{Y}_{\text{mod}}$ and measured $\mathbf{Y}_{\text{exp}}$ cluster concentrations:

$$F(\boldsymbol{\theta}_{\text{new}}) = \sum_{i=1}^{N} \frac{(Y_{\text{exp,i}} - Y_{\text{mod,i}}(\boldsymbol{\theta}_{\text{new}}))^2}{\sigma_i^2}, \tag{1}$$

where N stands for the number of measurements in synthetic data. We consider two sets of synthetic cluster concentrations: time-dependent, measured at $T = 278$ K and steady-state, measured for two temperatures (at $T = 278$ K and $T = 292$ K), as explained in Section 2.1. For the time-dependent synthetic data $N = N_C \times N_t$, where $N_C = 16$ stands for the number of cluster types included into simulations, while $N_t = 41$ stands for the number of time-step measurements available for each of the cluster types. For the second data set, $N = N_C \times N_T$, where $N_T = 2$ denotes the number of experiments conducted at different temperatures. In the formula above we scale the squared residuals by the measurement error variance σ_i^2 to avoid overfitting to the larger concentration values. The error variance σ_i^2 is matched depending on cluster type, time instance and temperature. See A2 for more details.

At each iteration of the Metropolis algorithm, the value $F(\boldsymbol{\theta}_{\text{new}})$ is compared to the least-square sum from the previous step $F(\boldsymbol{\theta}_{\text{old}})$. If the new value is lower (i.e., the candidate parameters fit the data at least as good as the the old values), then the step is accepted. In the opposite case, when $F(\boldsymbol{\theta}_{\text{new}}) > F(\boldsymbol{\theta}_{\text{old}})$, the point will be accepted with the probability

$$\alpha_{\text{acc}} = \exp\left[-\frac{1}{2}(F(\boldsymbol{\theta}_{\text{new}}) - F(\boldsymbol{\theta}_{\text{old}}))\right]. \tag{2}$$

If the candidate point is accepted, the parameter combination $\boldsymbol{\theta}_{\text{new}}$ is added to the chain, in the opposite case the old value is replicated in the chain. Finally, the value $F(\boldsymbol{\theta}_{\text{old}})$ is replaced with $F(\boldsymbol{\theta}_{\text{new}})$ and saved for the next iteration."

In this paper we employ a variant of the Metropolis algorithm which is more efficient at parameter sampling when the parameter space is large (Haario (2006)). This variant is called the Delayed Rejection Adaptive Metropolis (DRAM), introduced in Haario (2006). We briefly explain our approach below.

- We move the text starting from the line 134 ("We remark that to create a reliable sample from the underlying parameter distribution..") and ending at the end of the paragraph to Section 2.2.3 ("The overview of the MCMC runs").
- We move the lines 142-143 to the end of the Section 2.1.
- In line 142 we insert the Section 2.2.2 "The DRAM algorithm".
- In line 144 we add the sentence to "Similar to the basic Metroplois algorithm, the DRAM is initialized with the prior distribution and the initial guess for parameter values."
- In line 150, we cut the word "predefined".
- We move the Tables 3 and 4 to Section 2.2.3, titled as "The overview of the MCMC runs".
- We move the lines 143-144 to the end of the Section 2.2.2. We insert them after the description of the DRAM algorithm (after the line 188).
- We move the explanations of prior limits used for sampling the evaporation rates and thermodynamic data (lines 147-154) to Section 2.2.3.
- Starting from line 154, we changed the paragraph to "We make our initial guess $\boldsymbol{\theta} = \boldsymbol{\theta}_{old}$, where the prior distribution is flat; i.e., all the values within the upper and lower limits that were chosen for the sampled parameters are equally probable. The limits are summarized in Table 4. We also assume that the conditional probability distributions for the parameters given the concentration data are of Gaussian type.

  Once initialized, the following iterative steps take place. From the previous point in the MCMC chain $\boldsymbol{\theta}_{old}$, a new candidate for the unknown parameter values, $\boldsymbol{\theta}_{new}$, is sampled using the Gaussian proposal distribution. We then use the algorithm in Section 2.1

to obtain concentration outputs from the evaporation rates $\boldsymbol{\theta}_{new}$. In the first stage of DRAM, we chose to accept the new proposed values $\boldsymbol{\theta}_{new}$ with probability ... "

R:

- Changed in line 162 "... the concentrations obtained from the ACDC and VODE simulations with parameters $\boldsymbol{\theta}_{old}$ and $\boldsymbol{\theta}_{new}$, respectively."
- After the paragraph 186-189 we insert the Section 2.2.3 with the caption "The overview of the MCMC runs".
- At the beginning of the Section 2.2.3 we insert the paragraph "In our implementation of the DRAM algorithm, we impose upper and lower limits for the parameter values. We add such domain restrictions to exclude unphysical estimates for our parameters. These restrictions are encoded in our prior distribution, which we set to be a combination of so-called "flat priors", which are distributions that are proportional to a constant, (see Tables 3-4)."
- Next, we include an explanation of the prior distribution and physical limitations for the sampled parameters, which starts as follows: "We emphasize that there are currently no theoretical principles or experimental results which indicate possible restrictions for even the order of magnitude of the evaporation rates."
- After the domain restrictions, we explain the parameterization that we use for the evaporation rates and illustrate the sampling procedure (with Figure 1), i.e., we insert the lines 191-218.
- Next we insert the lines 134-138, starting from the sentence "We remark that to create a reliable sample from the underlying 135 parameter distribution...".
- We conclude the Section 2.2.3 with the lines 132-134, where we rephrase the sentences: "In all simulations of the algorithm given in the previous section, the sets of parameters which produce cluster concentrations within the allotted noise level of the data are kept in the chain. More specifically, the sampled parameters 270 of the posterior distribution represent the model evaluations which produce values within the noise level of 0.001% of the data concentrations for each of the respective cluster types".

2. **C:** It is quite confused that throughout the paper, the authors use identification of the rates and thermodynamic enthalpies/entropies. Is it

better to use for example estimate or similar words?

R: It is common language to use the words "identification/identify/determine/etc." in the inverse problems literature. We have changed some instances of these words to "estimate/estimation" to suit the atmospheric audience.

3. **C:** For pairwise marginal posterior distributions, either for evaporation rates or enthalpies/entropies, what criteria the authors used to create these correlations? For example, it seems that evaporation of different monomers from different clusters might be irrelevant.

   R: We created pairwise marginal posterior distributions from the history of the sampled chains for both cases: in case of evaporation rates and thermodynamic parameters. We observe that the evaporations of different monomers are correlated for some of the cluster types. For example, see Figure C4 and the monomer evaporations from $\left(H_2SO_4\right)_2\left(NH_3\right)_1$; and Figure C7 and the monomer evaporations from $\left(H_2SO_4\right)_5\left(NH_3\right)_4$ which display non-linear correlations. Also the evaporation rates for different non-monomers from different clusters can be correlated. For example, see Figure C7, where the evaporation rates $\left(H_2SO_4\right)_4\left(NH_3\right)_4 \rightarrow \left(H_2SO_4\right)\left(NH_3\right)$ and $\left(H_2SO_4\right)_5\left(NH_3\right)_3 \rightarrow \left(H_2SO_4\right)_2\left(NH_3\right)$ that display inverse linear correlation. However, as the reviewer had mentioned, the evaporation of different monomers from different clusters is irrelevant.

4. **C:** Section 3.4: can the authors present more details of the comparison instead of just some dry descriptions? For example, the authors can add a table to summarize the knowledge up-to-date regarding the evaporation rates from both measurements and modeling so that the readers can be benefit from reading this paper.

   C: We add a short summary paragraph regarding the evaporation rates and how they can be obtained: "The evaporation rates can be obtained either experimentally or computationally, when applying the Quantum Chemical (QC) methods, see Kürten, 2019. Experimental detection was conducted from the measurements in a flow tube (Hanson and Eisele, 2002; Jen et al., 2014, 2016; Hanson et al., 2017) and in the CLOUD chamber (Kurtén et al., 2007; Nadykto and Yu, 2007; Ortega et al., 2012; Elm et al., 2013; Elm and Kristensen, 2017; Yu et al., 2018). However, experimental detection is only available for the charged clusters. The summary of thermodynamic parameters obtained from different methods has previously been published in Kürten, 2019. These parameters can be employed to calculated the evaporation

rates at different temperatures."

5. **C:** Can the authors give some plausible explanation why evaporation rates estimated from transient data seem better than those from steady-state data?

   R: The transient data is a larger data set than that of just the steady-state data at one temperature. The extra information contained in the transient data reduces the size of the space of allowable evaporation rates, as it there are more restrictions on the possible values the evaporation rates make take. Also the transient data contain information about the slope of the concentrations changing with time, which contributes to quantification of the associated processes (such as collisions and evaporations). We have added the following sentences to emphasize this point:

   - Starting in line 262, we change the paragraph to " First, we extend the synthetic measurement data from steady state concentrations to transient concentrations. The data set for transient cluster concentrations at one temperature is larger than the data set for steady-state cluster concentrations at one temperature, as the transient data contains the concentration values at multiple times instances. Also the transient data contain information about the slope of the concentrations changing with time (see Figure C1), which contributes to quantification of the molecular-scale processes (such as collisions and evaporations). We thus expect that this larger data set will reduce the dimension of the solution space for the evaporation rates. Indeed, we will show that this is the case. We generate a synthetic transient cluster concentration data set using the method in Section 2.1. The time resolution of our new synthetic data set is 1.5 minutes, which results in ~~2624~~ 656 total concentration measurements for all the cluster type measured for four different ammonia concentrations. These data sets are illustrated in Figure C1. "

     Then in line 267, we added: "From this transient cluster concentration data set, we then conduct analogous MCMC runs (as described in Section 2.2). As in the steady-state ..."

   - Here we summarize the main differences between the steady-state and transient data as follows: "In the case of the steady-state cluster concentrations we include only one value for each of the 16 cluster types considered in the study, which were taken when

the system has attained a steady state (at the end of the ACDC simulation). The transient data contain the steady-state data as subset. Specifically, in this case we consider the concentrations measured when the system has attained the steady state together with the time-step concentration data measured from the starting point to the end of the ACDC simulation."

6. **C:** The authors claimed that the 5A5N has low variance in free energies. However, an order of magnitude is not small for free energies and it is substantial if this value is applied to the evaporation rates (Line 319 on p18).

   R: We change the sentence in line 319 to: "Although the posterior distributions of sampled thermodynamic parameters for $\left(H_2SO_4\right)_5\left(NH_3\right)_5$ feature higher uncertainties in comparison to the corresponding posterior distributions identified for the smaller clusters, the evaporation rates for evaporations from $\left(H_2SO_4\right)_5\left(NH_3\right)_5$, as calculated from the aforementioned posterior distributions, have low variances, see Table D3."

   R: Note to TB: We will rather point out that the evaporation rates for the biggest cluster calculated from a posterior distributions of thermodynamic parameters feature low variances. Do you agree?

7. **C:** There are several rather minor comments below:

   (a) P11, lines 233, do the authors mean that the lower limits of evaporation of a monomer from those clusters are far above the 10ˆ-10 as defined for complete growth?

   (b) P11, line 240, Figures 3-4 can actually be combined to one figure since they basically represent different parts of the same thing. There are some figures that have similar issues.

   (c) P15, Figure 5, no label for a, b, c, d.

   (d) P15, line 284, how the evaporation rates of monomers for clusters 2A display inverse linear correlations in Figures C4-C8?

   (e) P18, the claim that the estimated formation enthalpies vary at most by 1 kcal mol^{-1}, while the variance for the formation entropies is less than 1 $calK^{-1}mol^{-1}$ is not right.

   R: We calculated the variances of estimated parameters and the claim will be corrected by replacing the sentence in P18 with "It can be seen that for all the clusters except $\left(H_2SO_4\right)_5\left(NH_3\right)_5$ the

variance for the estimated formation enthalpies are less than 0.46 kcal mol^{-1}, while the estimated formation entropies vary at most by 5.4 cal $K^{-1}mol^{-1}$."

(f) P18, line 313 and line 321, Figure 9 should not appear before figure 8.

(g) There are lot of typos of molecular sulfuric acid formula throughout the manuscript and a thorough check should be made before submitting the revision. For example, H2SO2.

(h) The references cited in the text are not followed the journal guidelines.

(i) Line 34 on p2, subscript; line 37, miss a comma? Line 39, "," is surplus.

(j) Line 54 on p3, "-" superscript? line 59, miss a comma between experiment and these? It is apparent an ill-sentence (line 65).

(k) Line 104 on p4, into instead of in to?

(l) Table 1, it is suggested to add a third column to indicate the number of clusters in each row.

(m) Line 123 on p5, kinetic model?

(n) Line 369 on p23, what is question mark for?

(o) Figure D2, kkal/mol?

R: We have made changes to the document to correct for these typos. We are very grateful to the the referee for their careful eye!